Video Genre identification by spoken content analysis

Video indexing is a challenging task in a global context of fast growing of digital TVs and video collections on the Internet. Structuring such collections requires high level categorization of contents.

Here, we propose to identify the genre of a video by analysing the audio contents. Video genre refers to the editorial style of the video. In these experiments, we consider 7 genres : commercials, news, cartoons, music, documentary, movies and sports.

The targeted task consists in recognizing the genre by processing only the audio channel, without any video features. Your Genre Identification (GID) system will contain two modules. The first one performs the feature extraction, and the second one classifies the feature vectors.

Feature extraction

Here we consider features extracted from the outputs of an Automatic Speech Recognition System, including speaker turns, speaker tracking and automatic transcriptions.

You can get automatic transcription of the video contents at http://filez.univ-avignon.fr/c5q. It is important to note that WEB data are especially difficult to recognize with a classical ASR system, due to content and context diversity.

1/ Identification by meaningful words

A popular approach of text categorisation consists in using vectorial representation of documents : each document is represented by a vector of word frequencies. Assuming that meaningful words are more relevant than tool-words, many authors proposed to remove the most frequent word of the language (a stoplist can be found at http://sites.univ-provence.fr/veronis/data/antidico.txt). You can also estimate your own stoplist on the trasncripts.

You have to test this approach by extracting statistics from the provided automatic transcriptions and to train a SVM classifier by using libsvm toolbox (a brief tuytorial can be found at http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf).

2/ Identification by tool-words only.

Here, the basic idea is that frequencies of these stopwords are characteristic of the video genre. Unlike the classical TF-IDF approach, the proposed method is topic-independent, stopwords being not topic-related.

Moreover, the automatic transcription is obtained with an LM that is not adapted to the documents; the words that are out of the lexicon will be missed. Thus, the transcriptions contain errors. We assume 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 Vs with the *n* most frequent words in the transcriptions of the training corpus.

Classification

Classification will be achieved by using a Support Vector Machine (SVM) classifier. Training must be conducted on the working corpus available. It includes a training set composed of 1800 videos and a test set composed of 200 videos.

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

[2] Mickael Rouvier, Georges Linarès, and Driss Matrouf, “Robust audio-based classification of video genre,” in Proc. INTERSPEECH , 2009, pp. 1159–1162.

[3] W.H. Lin and A. Hauptmann, “News video classification using svm-based multimodal classifiers and combination strategies,” in Proc. ICM , 2002, pp. 323–326.

[4] D. Brezeale and D.J. Cook, “Using closed captions and visual features to classify movies by genre,” in Proc. MDM/KDD , 2006.

[5] W. Qi, L. Gu, H. Jiang, X.R. Chen, and H.J. Zhang, “Integrating visual, audio and text analysis for news video,” in Proc. ICIP , 2000, vol. 3.
[6] T. Tokunaga and I. Makoto, “Text categorization based on weighted inverse document frequency,” SIG-IPSJ, pp. 33–39, 1994.
[7] C.D. Manning, P. Raghavan, and H. Sch¨utze, Introduction to Information Retrieval, Cambridge University Press, 2008.
[8] G. Forman, “An extensive empirical study of feature selection metrics for text classification,” The Journal of Machine Learning Research , vol. 3, pp. 1289–1305, 2003.
[9] M. Roach and J. Mason, “Classification of video genre using audio,” in Proc. ECSCT , 2001.
[10] G. Linar`es, P. Noc´era, D. Massonie, and D. Matrouf, “The lia speech recognition system: from 10xrt to 1xrt,” in Lecture Notes in Computer Science , 2007.
[11] Mickael Rouvier, Driss Matrouf, and Georges Linarès, “Factor analysis for audio-based video genre classification,” in Proc. INTERSPEECH , 2009, pp. 1155–1158.
[12] J. Ajmera, I. McCowan, and H. Bourlard, “Robust HMM-based speech/music segmentation,” in Proc. ICASSP , 2002, vol. 1, pp. 297–300.