KIT at MediaEval 2011 - Content-based genre classification on web-videos

Tomas Semela
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT)
76131 Karlsruhe, Germany
tomas.semela@student.kit.edu

Hazım Kemal Ekenel
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT)
76131 Karlsruhe, Germany
ekenel@kit.edu

ABSTRACT

In this paper, we run our content-based video genre classification system on the MediaEval evaluation corpus. Our system is based on several low level audio-visual cues, as well as cognitive and structural information. The purpose of this evaluation is to assess our content-based system's performance on the diversified content of the blip.tv web-video corpus, which is described in detail in [5].

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

Keywords

Genre classification, content-based features

1. MOTIVATION

Automatic genre classification is an important task in multimedia indexing. Several studies have been conducted on this topic. A comprehensive overview of these studies on TV genre classification can be found in [6]. Recently, there has also been an increasing interest in web video genre classification [9][1]. In this study, we evaluated our content-based system, which is based on the low-level audio-visual features, on the MediaEval corpus. The utilized features in the system correspond to low level color and texture cues, as well as shot boundary and face detection outputs. We used this features before for detecting high-level features in videos [2] and successfully classified various TV content into genres [3]. In the following sections we give a brief overview of our system, for details please refer to [3].

2. CONTENT-BASED FEATURES

2.1 Cognitive and structural features

Cognitive and structural features are proposed in [6]. Cognitive features are derived using a face detector. It contains average number of faces per frame, distribution of number of faces per frame and distribution of location of the faces in the frame. Structural feature is derived using a shot boundary detector. It contains average shot duration and distribution of shot lengths. The cognitive and later presented visual features are extracted from 5 linearly distributed frames per shot.

2.2 Aural Features

To benefit from the audio information of each clip, we compute four features from the audio signal. All features are extracted from mono-channel audio with 16 kHz sample rate and a 256 kbit/s bit rate. The features include *MFCC*, *Zero Crossing Rate* and *Signal Energy*, and are utilized using different representations.

2.3 Low-level Visual Features

We used six different low level visual features which represent color and texture information in the video.

2.3.1 Color descriptors

Histogram: We use the HSV color space and build a histogram with 162 bins [8].

Color moments: We use a grid size of 5×5. The first three order color moments were calculated in each local block in the image and the Lab color space is used [7].

Autocorrelogram: Autocorrelogram captures the spatial correlation between identical colors. 64 quantized color bins and five distances are used [4].

2.3.2 Texture descriptors

Co-occurrence texture: As proposed in [1], five types of features are extracted from the gray level co-occurrence matrix (GLCM): Entropy, Energy, Contrast, Correlation and Local homogeneity.

Wavelet texture grid: We calculate the variances of the high-frequency sub-bands of the wavelet transform of each grid region. We performed 4-level analysis on a grid that has $4 \times 4 = 16$ regions. Haar wavelet is employed, as in [1].

Edge histogram: For the edge histogram, 5 filters as proposed in the MPEG-7 standard are used to extract the kind of edge in each region of 2×2 pixels. Then, those small regions are grouped in a certain number of areas (4 rows $\times$ 4 columns in our case) and the number of edges matched by each filter (vertical, horizontal, diagonal 45°, diagonal 135° and non-directional) are counted in the region's histogram.

3. CLASSIFICATION

Classification is performed using multiple SVM classifiers. As can be seen in Fig. 1, content-based features are extracted from each video and are used as input for separate

[1]Also as part of the ACM Multimedia Grand Challenge

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

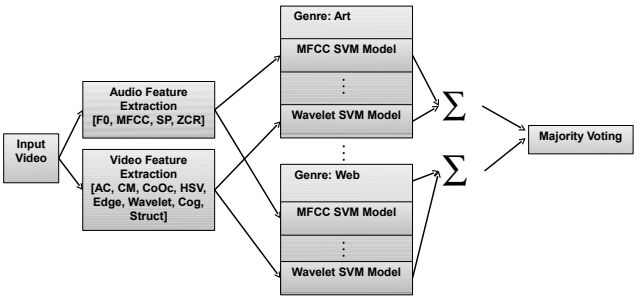

Figure 1: System Overview

SVMs, one for each genre and feature. Classification output of each SVM is summed up over all features for each genre and a genre is picked via majority voting.

4. EVALUATION AND DISCUSSION

The evaluation of this years MediaEval genre tagging task was performed on 1727 clips form blip.tv, distributed unevenly over 26 categories including a *default* category. Single label classification is performed and mean average precision (MAP) is used as the official performance measure. Training of the SVMs was conducted on approximately 100 videos for each genre, except for autos_and_vehicles where only 14 clips were available. These training videos are from a larger additional set of blip.tv videos. However, since we had limited time, it was not possible to process all these videos. Therefore, we limited the number of training videos per genre to 100 videos, which are randomly selected for each genre. Because our system works as a single label classification system, we also computed simple classification accuracy and calculated a 2nd MAP performance with a similarity value of 1, instead of a very low probability output of our system.

All in all, 5 runs were evaluated using these three evaluation measures. In our case, a combination of all feature sets (run1) and each feature category like visual (run2), aural (run3), cognitive (run4) and structural (run5) are evaluated independently. The results are presented in Table 1. The least contribution comes from the cognitive features, while the visual features (run 2) contribute the most to the overall performance, outperforming the other runs in the MAP performance measures and achieving almost the same classification accuracy as all feature sets together. From the six available visual features *color moments* and *wavelet texture* show the best classification results with 20% and 23%, respectively.

The best results (greater 50%) were achieved in the web_development (66.6%), mainstream_media (68.9%), food_and_drink (61.1%), movies_and_television (58.5%) and literature category with 89.6%. Worst results (under 10%) showed documentary (4.5%), educational (3.2%), health (9.5%), travel (7.1%) and videoblogging with 0%.

	run1	run2	run3	run4	run5
MAP	0.0023	0.0035	0.001	0.001	0.003
2nd MAP	0.0038	0.006	0.001	0.0012	0.0028
Accuracy (%)	28.2	27.5	13.9	1.3	5.4

Table 1: Evaluation Results

Our experiments show that a content-based system which is able to achieve nearly perfect accuracy on TV datasets (95% and 99%, see [3]) and also very high performance on a YouTube dataset (92.4%), is not able to achieve high performance on the blip.tv corpus. The main reason for this might be the increased number of genres to be classified, high intra-class diversity leading to difficulty in seperability of genres from each other using content-based cues.

More interestingly the low-level visual and aural features show more promising results than the selected higher-level cognitive and structural cues. Either it is not possible to cover the variety, or overall resemblance of all videos with these features or more promising high-level features have to be found by analyzing the properties of the web-videos.

Because of the limits of content-based systems in this area, the usage of metadata and other sources like ASR engines is desirable to be able to attain a robust genre classification system.

Acknowledgments

This study is funded by OSEO, French State agency for innovation, as part of the Quaero Programme.

5. REFERENCES

[1] M. Campbell, E. Haubold, S. Ebadollahi, D. Joshi, M. R. Naphade, A. P. Natsev, J. Seidl, J. R. Smith, K. Scheinberg, and L. Xie. IBM Research TRECVID-2006 Video Retrieval System. In *Proc. of NIST TRECVID Workshop 2006*.

[2] H. K. Ekenel, H. Gao, and R. Stiefelhagen. Universität Karlsruhe (TH) at TRECVID 2008. In *NIST TRECVID Workshop*, Gaithersburg, USA, Nov. 2008.

[3] H. K. Ekenel, T. Semela, and R. Stiefelhagen. Content-based video genre classification using multiple cues. In *Proceedings of the 3rd International Workshop on Automated Information Extraction in Media Production*, AIEMPro'10, pages 21–26, 2010.

[4] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. Image indexing using color correlograms. In *Computer Vision and Pattern Recognition (CVPR)*, pages 762–768, 1997.

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

[6] M. Montagnuolo and A. Messina. Parallel neural networks for multimodal video genre classification. *Multimedia Tools Appl.*, 41:125–159, January 2009.

[7] M. A. Stricker and M. Orengo. Similarity of color images. In *Storage and Retrieval for Image and Video Databases (SPIE)'95*, pages 381–392, 1995.

[8] M. J. Swain and D. H. Ballard. Color indexing. *International Journal of Computer Vision*, 7:11–32, 1991.

[9] Z. Wang, M. Zhao, Y. Song, S. Kumar, and B. Li. YouTubeCat: Learning to categorize wild web videos. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 879 –886, June 2010.