KIT at MediaEval 2012 – Content–based Genre Classification with Visual Cues

Tomas Semela
tomas.semela@student.kit.edu

Makarand Tapaswi
makarand.tapaswi@kit.edu

Hazım Kemal Ekenel
ekenel@kit.edu

Rainer Stiefelhagen
rainer.stiefelhagen@kit.edu

Institute for Anthropomatics, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany

ABSTRACT

This paper presents the results of our content–based video genre classification system on the 2012 MediaEval Tagging Task. Our system utilizes several low–level visual cues to achieve this task. The purpose of this evaluation is to assess our content–based system's performance on the large amount of *blip.tv* web–videos and high number of genres. The task and corpus are described in detail in [5].

Categories and Subject Descriptors

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

General Terms

Content, video, genre, classification

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 on TV genre classification can be found in [4]. Recently, there has also been an increasing interest in web video genre classification [9][1]. In this year's study, we evaluated new additions to our content–based system on the MediaEval corpus. The utilized features in the system correspond to low–level color and texture cues as well as newly added SIFT descriptors. We used the color and texture cues before to successfully classify various TV content into genres [2]. In the following sections we give a brief overview of our system, for details please refer to [2].

2. CONTENT–BASED FEATURES

2.1 Low-level Visual Features

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

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

MediaEval 2012 Workshop, October 4-5, 2012, Pisa, Italy

2.1.1 Color descriptors

Histogram: We use the HSV color space and build a histogram with 162 bins [7].

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 [6].

Autocorrelogram: Autocorrelogram captures the spatial correlation between identical colors. 64 quantized color bins and five distances are used [3].

2.1.2 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.

2.2 Bag–of–words

SIFT features in combination with the bag–of–words model proved to be very successful in computer vision over the past few years. For e.g., authors in [8] use it to perform scene recognition. As part of our system, we extract rgb-SIFT features from every keyframe of each video computed with the dense sampling point selection strategy. The final feature vector of each keyframe is 500–dimensional. The codebook is computed using k–Means and 1500 keyframes from the development set evenly distributed over all genres.

3. CLASSIFICATION

Classification is performed using multiple SVM classifiers. As can be seen in Fig. 1, content–based features are extracted from the provided keyframes of the corpus and are used as input for separate SVMs, one for each genre and feature. The data from the development set was used to train these SVM classifiers. Classification output of each SVM is summed up over all features for each genre and a genre is picked via majority voting. In case of color and texture

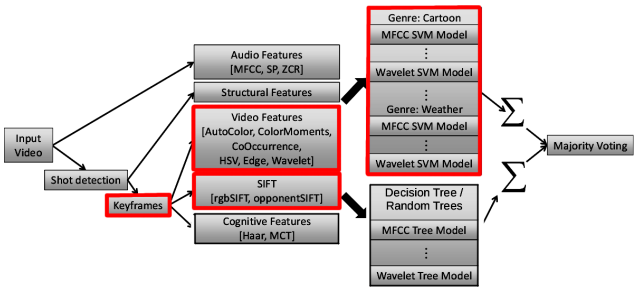

Figure 1: System Overview: Boxes marked in red are used

cues, one feature vector per video is used as input. For SIFT classification, each keyframe is classified separately and predictions for each video are averaged over all keyframes. This average is treated as another single feature SVM output.

Prior domain knowledge can be included optionally in the classification process. The prior domain knowledge is the video distribution over the genres in the development set. For example the *autos_and_vehicles* genre contains only 8 videos while *politics* and *technology* are represented with around 500 videos. This is not only a bias of the dataset, but a bias of the categories themselves, as one could consider that *technology* incorporates *autos_and_vehicles*. We thus use this information in the final prediction of the genre class as a likelihood of the genre distribution on *blip.tv*. We first accumulate genre scores over all feature SVM outputs. These SVM scores for each video are then normalized to unit sum. We then divide these probabilities by the square root of the number of videos in the development set for each genre to include the a–priori knowledge of the class distribution. Finally, step one is repeated to obtain unit sum.

4. EVALUATION

The evaluation of this year's MediaEval genre tagging task includes 9550 clips from *blip.tv*, distributed over 26 categories including a *default* category. Only 9457 videos came with provided keyframes and since our system is visual–based, only these videos were tagged. Single label classification is performed and mean average precision (MAP) is used as the official performance measure.

We evaluated 6 different runs that are not part of the 5 official runs, since our system is visual content–based only. The results are presented in Table 1. Run_1 and run_4 utilize only color and texture descriptors. Run_2 and run_5 evaluate the rgbSIFT feature and finally run_3 and run_6 evaluate the combination of both. The runs are divided into two groups where the same set of features are evaluated with and without prior domain knowledge.

Features	Color Texture -	- - rgbSIFT	Color Texture rgbSIFT
No Knowledge	run_1	run_2	run_3
MAP	**0.3008**	**0.2329**	**0.3499**
Prior Knowledge	run_4	run_5	run_6
MAP	**0.3461**	**0.1448**	**0.3581**

Table 1: Evaluation Results

Run_6 reaches a MAP score of 0.358 using a–priori domain knowledge and all available visual features. The rgbSIFT feature alone performed poorly compared to the other runs and is the only case where prior domain knowledge worsens the results with a final MAP score of 0.144. In general the domain knowledge seems to improve the classification.

Looking at the top and worst AP scores for individual genres in run_6, the results show that the genres *autos_and_vehicles* (0.812), *health* (0.668), *movies_and_television* (0.602), *religion* (0.578) and *food_and_drink* (0.566) achieve best results. The genres *travel* (0.010), *videoblogging* (0.100), *documentary* (0.119) and *citizen_journalism* (0.158) are most difficult to classify.

5. CONCLUSIONS

Compared to our results from last year's evaluation, the most recognizable and difficult genres stay mainly the same, e.g., *food_and_drink*, *movies_and_television*, *documentary*, *travel* and *videoblogging*. The only exception is the *health* genre, which switched from one of the most difficult genres to the top 5 best results.

Owing to the limitation of content–based visual information in the web video domain and the high number of genres, the usage of metadata and other sources like ASR transcripts is desirable to improve the results. Especially the more difficult genres, as shown in the evaluation, should benefit greatly from information from a different source.

Acknowledgments

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

6. 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*, Gaithersburg, USA, 2006.

[2] 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.

[3] 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.

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

[5] S. Schmeideke, C. Kofler, , and I. Ferrane. Overview of MediaEval 2012 Genre Tagging Task. In *MediaEval 2012 Workshop*, Pisa, Italy, October 4-5 2012.

[6] 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.

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

[8] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Empowering Visual Categorization with the GPU. *IEEE Transactions on Multimedia*, 13(1):60–70, 2011.

[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.