Audio-Visual Content Description for Video Genre Classification in the Context of Social Media

Bogdan Ionescu[1,3], Klaus Seyerlehner[2], Constantin Vertan[1], Patrick Lambert[3]
[1]LAPI - University Politehnica of Bucharest, 061071 Bucharest, Romania
{bionescu,cvertan}@alpha.imag.pub.ro
[2]DCP - Johannes Kepler University, A-4040, Linz, Austria
klaus.seyerlehner@gmail.com
[3]LISTIC - Polytech Annecy-Chambery, B.P. 80439, 74944 France
patrick.lambert@univ-savoie.fr

ABSTRACT

In this paper we address the automatic video genre classification with descriptors extracted from both, audio (block-based features) and visual (color and temporal based) modalities. Tests performed on 26 genres from blip.tv media platform prove the potential of these descriptors to this task.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*audio, color and action descriptors*; I.5.3 [**Pattern Recognition**]: Clustering—*video genre*.

Keywords

block-based audio features, color perception, action content, video genre classification.

1. INTRODUCTION

In this paper we address the issue of automatic video genre classification in the context of social media platforms as part of the MediaEval 2011 Benchmarking Initiative for Multimedia Evaluation (see $http://www.multimediaeval.org/$). The challenge is to provide solutions for distinguishing between up to 26 common genres, like "art", "autos", "business", "comedy", "food and drink", "gaming", and so on [2]. Validation is to be carried out on video footage from the blip.tv media platform (see $http://blip.tv/$).

We approach this task, globally, from the classification point of view and focus on the feature extraction step. For a state-of-the-art of the literature see [1]. In our approach, we extract information from both audio and visual modalities. Whether these sources of information have been already exploited to genre classification, the novelty of our approach is in the content descriptors we use.

2. VIDEO CONTENT DESCRIPTION

Audio descriptors. Most of the common video genres tend to have very specific audio signatures, e.g. music clips contain music, in sports there is the specific crowd noise, etc. To address this specificity, we propose audio descriptors which are related to rhythm, timbre, onset strength, noisiness and vocal aspects. The proposed audio features are block-level based, which compared to classic approaches have the advantage of capturing local temporal information by analyzing sequences of consecutive frames in a time-frequency representation. Audio information is described with parameters such as: *spectral pattern* (characterize the soundtrack's timbre), *delta spectral pattern* (captures the strength of onsets), *variance delta spectral pattern* (captures the variation of the onset strength over time), *logarithmic fluctuation pattern* (captures the rhythmic aspects), *spectral contrast pattern* (estimates *"tone-ness"*) and *correlation pattern* (captures the temporal relation of loudness changes over different frequency bands). For more information see [3].

Temporal descriptors. The genre specificity is reflected also at temporal level, e.g. music clips tend to have a high visual tempo, documentaries have a reduced action content, etc. To address those aspects we detect sharp transitions, cuts and two of the most frequent gradual transitions, fades and dissolves. Based on this information, we assess *rhythm* as movie's average shot change speed computed over $5s$ time windows (provides information about the movie's changing tempo) and *action* in terms of high action ratio (e.g. fast changes, fast motion, visual effects, etc.) and low action ratio (the occurrence of static scenes). Action level is determined based on user ground truth [4].

Color descriptors. Finally, many genres have specific color palettes, e.g. sports tend to have predominant hues, indoor scenes have different lighting conditions than outdoor scenes, etc. We assess color perception by projecting colors onto a color naming system (associating names with colors allows everyone to create a mental image of a given color or color mixture). We compute a *global weighted color histogram* (movie's color distribution), an *elementary color histogram* (distribution of basic hues), *light/dark*, *saturated/weak-saturated*, *warm/cold* color ratios, *color variation* (the amount of different colors in the movie), *color diversity* (the amount of different hues) and *adjacency/complementarity color ratios*. For more information on visual descriptors see [4].

3. EXPERIMENTAL RESULTS

Results on development data. First validation was performed on the provided development data set (247 sequences) which was eventually extended to up to 648 sequences in order to provide a consistent training data set for classification (source blip.tv; sequences are different than the ones proposed for the official runs). We observed that in the case

Copyright is held by the author/owner(s).
MediaEval 2011 Workshop, September 1-2, 2011, Pisa, Italy

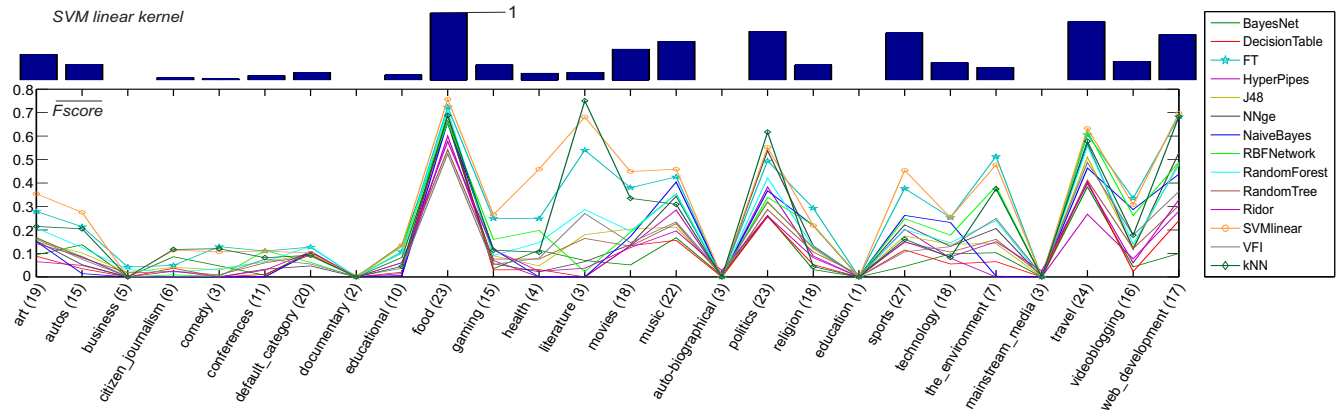

Figure 1: Average $Fscore$ achieved using all audio-visual descriptors and genre classification "performance" for the best run, i.e. SVM with linear kernel (graph on top).

of some of the proposed genres, the genre specific content is captured mainly with the textual information. Therefore, our tests focused mainly on genres with specific audio-visual contents, like "art", "food", "cars", "sports", etc. (for which we provided a representative number of examples).

Tests were performed using a cross-validation approach. We use for training p% of the existing sequences (randomly selected and uniformly distributed with respect to genre) and the remainder for testing. Experiments were repeated for different combinations between training and testing (e.g. 1000 repetitions).

Figure 1 presents average $\overline{F_{score}} = 2 \cdot \overline{P} \cdot \overline{R}/(\overline{P} + \overline{R})$ ratio (where $\overline{P}$ and $\overline{R}$ are average precision and recall, respectively over all repetitions) for $p = 50\%$, descriptor set = audio-color-action (i.e. the descriptor set which provided the most accurate results) and various classification approaches (see Weka at $http://www.cs.waikato.ac.nz/ml/weka/$). The number in the brackets represent the number of test sequences used for each genre.

From the global point of view, the best results are obtained with SVM and a linear kernel (depicted in Orange), followed by k-NN ($k = 3$, depicted in Dark Green) and FT (Functional Trees, depicted in Cyan). At genre level, the best accuracy is obtained for genres with particular audio-visual signatures. The graph on top presents a measure of the individual genre classification "performance" which is computed as the $\overline{F_{score}}$ times the number of test sequences used. An $\overline{F_{score}}$ obtained for a greater number of sequences is more representative than one obtained for only a few (values are normalized with respect to 1 for visualization purpose). The proposed descriptors provided good discriminative power for genres like (the number in the brackets is $\overline{F_{score}}$): "food and drink" (0.757), "travel" (0.633), "politics" (0.552), "web development and sites" (0.697), while at the bottom end are genres whose contents are less reflected with audio-visual information, e.g. "citizen journalism", "business", "comedy" (see Figure 1).

Results on test data. For the final official runs, classification was performed on 1727 sequences with training performed on the previous data set (648 sequences). The overall results obtained in terms of MAP (Mean Average Precision) are less accurate than the previous results, thus: 0.077 for k-NN on audio-color-action, 0.027 for RandomForest on audio-color-action, 0.121 **for SVM linear on audio-color-action** (best run), 0.103 for SVM linear on audio and 0.038 SVM linear on color-action. This is mainly due to the limited training data set compared to the diversity of test sequences and to the inclusion of the genres for which we obtain 0 precision (i.e. audio-visual information is not discriminant, see Figure 1). The fact that MAP provides only an overall average precision over all genres makes us unable to conclude on the genres which are better suited to be retrieved with audio-visual information and which fail to.

4. CONCLUSIONS AND FUTURE WORK

The proposed descriptors performed well for some of the genres, however to improve the classification performance a more consistent training database is required. Also, our approach is more suitable for classifying genre patterns from the global point of view, like episodes from a series being not able to detect a genre related content within a sequence. Future tests will consist on preforming cross-validation on all the 2375 sequences (development + test sets).

5. ACKNOWLEDGMENTS

Part of this work has been supported under the Financial Agreement EXCEL POSDRU/89/1.5/S/62557.

6. REFERENCES

[1] D. Brezeale, D.J. Cook, "Automatic Video Classification: A Survey of the Literature," IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 38(3), pp. 416-430, 2008.

[2] M. Larson, M. Eskevich, R. Ordelman, C. Kofler, S. Schmiedeke, G.J.F. Jones, "Overview of MediaEval 2011 Rich Speech Retrieval Task and Genre Tagging Task", MediaEval 2011 Workshop, Pisa, Italy, 2011.

[3] K. Seyerlehner, M. Schedl, T. Pohle, P. Knees, "Using Block-Level Features for Genre Classification, Tag Classification and Music Similarity Estimation," MIREX-10, Utrecht, Netherlands, 2010.

[4] B. Ionescu, C. Rasche, C. Vertan, P. Lambert, "A Contour-Color-Action Approach to Automatic Classification of Several Common Video Genres", AMR (LNCS 6817), Linz, Austria, 2010.