ARF @ MediaEval 2012: Multimodal Video Classification

Bogdan Ionescu[1,5]
bionescu@imag.pub.ro

Ionuţ Mironică[1]
imironica@imag.pub.ro

Klaus Seyerlehner[2]
music@cp.jku.at

Peter Knees[2]
peter.knees@jku.at

Jan Schlüter[3]
jan.schlueter@ofai.at

Markus Schedl[2]
markus.schedl@jku.at

Horia Cucu[4]
horia.cucu@upb.ro

Andi Buzo[4]
andi.buzo@upb.ro

Patrick Lambert[5]
patrick.lambert@univ-savoie.fr

ABSTRACT

In this paper we study the integration of various audio, visual and text-based descriptors for automatic video genre classification. Experimental validation is conducted on 26 video genres specific to web media platforms (e.g. blip.tv).

Categories and Subject Descriptors

H.3.1 [**Information Search and Retrieval**]: *video genre*.

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Multimodal video genre classification, early fusion.

1. INTRODUCTION AND APPROACH

We approach the automatic Genre Tagging Task @ MediaEval 2012 [1] from the perspective of machine learning techniques (we attempt to learn the specificity of each genre). To this end, we set the objective to test a broad range of classifiers and multimodal content descriptor combinations.

1.1 Audio descriptors

• **block-based audio** (11,242 values) - to capture the temporal properties of the audio signal, we propose a set of audio descriptors that are computed from overlapping audio blocks (= sequences of consecutive spectral frames). On each block we compute the *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 the *"tone-ness"*), *Correlation Pattern* (captures the temporal relation of loudness changes) and timbral features: *Local Single Gaussian Model* and *George Tzanetakis Model* of MFCCs (Mel-Frequency Cepstral Coefficients). Sequence aggregation is achieved by taking the mean, variance, or median over all blocks [2].

• **standard audio features** (196 values) - we also experimented with a set of general-purpose audio descriptors: *Linear Predictive Coefficients* (LPCs), *Line Spectral Pairs* (LSPs), *MFCCs*, *Zero-Crossing Rate* (ZCR), and *spectral centroid*, *flux*, *rolloff*, and *kurtosis*, augmented with the variance of each feature over a certain window (a common setup for capturing enough local context is 1.28 s). For a clip, we take the mean and standard deviation over all frames.

1.2 Visual descriptors

• **MPEG-7** (4,182 values) - we adopted some standard color and texture-based descriptors such as: *Local Binary Pattern* (LBP), *autocorrelogram*, *Color Coherence Vector* (CCV), *Color Layout Pattern* (CLD), *Edge Histogram* (EHD), *Scalable Color Descriptor* (SCD), *classic color histogram* (hist) and *color moments*. For each sequence, we aggregate the features by taking the mean, dispersion, skewness, kurtosis, median and root mean square statistics over all frames.

• **feature detectors** - from this category, we compute average *Histogram of oriented Gradients* (HoG) over all frames, *Speeded Up Robust Feature* (SURF) and the *Harris corner detector* via 4,000 word Bag-of-Visual-Words dictionaries.

1.3 Text descriptors

• **TF-IDF** (ASR-based 3,466 values, metadata-based 504) - we use the standard Term Frequency-Inverse Document Frequency approach. First, we filter the input text by removing XML markups and terms with a document frequency less than 5%-percentile of the frequency distribution. We reduce further the term space by keeping only those terms that discriminate best between genres according to the χ^2-test. We generate a global list by retaining for each genre class, the m terms (e.g. $m = 150$ for ASR and 20 for metadata) with the highest χ^2 values that occur more frequently than in complement classes. This results in a vector representation for each document that is subsequently *cosine normalized* to remove the influence of the length of transcripts.

2. EXPERIMENTAL RESULTS

We carry out a preliminary validation on the provided

[1] LAPI, University Politehnica of Bucharest, Romania (research grant EXCEL POSDRU/89/1.5/S/62557; we also thank Prof. Nicu Sebe, Univ. of Trento, for his support).
[2] DCP, Johannes Kepler University, Linz, Austria.
[3] Austrian Research Institute for Artificial Intelligence, Vienna, Austria.
[4] SpeeD, University Politehnica of Bucharest, Romania.
[5] LISTIC, Polytech Annecy-Chambery, France.

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

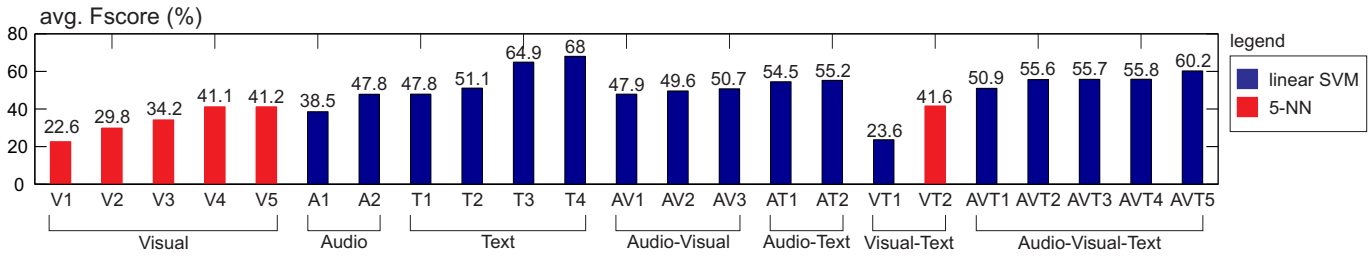

Figure 1: Genre average $Fscore = 2 \cdot P \cdot R/(P+R)$ ($P = precision$, $R = recall$; train-test percentage split of 50%).

V1	SURF with B-o-V-W
V2	Harris with B-o-V-W
V3	hist+HoG
V4	MPEG-7
V5	LBP+CCV+hist
A1	standard audio
A2	block-based audio[2]
T1	TF-IDF ASR LIUM[3]
T2	TF-IDF ASR LIMSI[4]
T3	TF-IDF metadata
T4	TF-IDF metadata+ASR LIMSI[4]
AV1	hist+HoG+block-based audio[2]
AV2	LBP+CCV+hist+block-based audio[2]
AV3	MPEG-7+block-based audio[2]
AT1	block-based audio[2]+TF-IDF ASR LIMSI[4]
AT2	block-based audio[2]+TF-IDF metadata
VT1	LBP+CCV+hist+TF-IDF metadata
VT2	LBP+CCV+hist+TF-IDF ASR LIMSI[4]
AVT1	standard audio+LBP+CCV+hist+TF-IDF ASR LIMSI[4]
AVT2	block-based audio[2]+LBP+CCV+hist+TF-IDF metadata
AVT3	block-based audio[2]+LBP+CCV+hist+TF-IDF ASR LIMSI[4]
AVT4	block-based[2]&standard audio+LBP+CCV+hist+TF-IDF ASR LIMSI[4]
AVT5	block-based[2]&standard audio+LBP+CCV+hist+TF-IDF metadata&ASR LIMSI[4]

devset (5,127 sequences) with the objective of determining the best descriptor-classifier combination (we experiment on Weka, `http://www.cs.waikato.ac.nz/ml/weka/` and using early fusion to combine descriptors). For the official runs, we train the classifier on the devset and use the provided *testset* [1] (9,550 sequences) for classification. In the sequel we present the best results that were achieved using k-Nearest Neighbor (k-NN) and Support Vector Machines (SVM).

2.1 Classification on devset

Figure 1 presents some of the classification results achieved on the devset. The first notable result is that (at least on this type of data - web video) the visual descriptors, regardless their nature (e.g. features, texture, etc.), are performing in an average $Fscore$ interval of 30%±10%. For instance, using all MPEG-7 descriptors is not more accurate than using only 3 of them (see V5 vs. V4). Therefore, computational power can be saved by using only few visual descriptors instead of using complex combinations (e.g. Bag-of-Visual-Words). The audio descriptors are performing slightly better than visual ones (highest improvement is up to 6%). The proposed block-based audio features [2] exceed the standard ones by at least 9%. Increasing the number of modalities tends to increase also the performance, e.g. by 12% to 19% (e.g. see AV3, AVT5). However, the highest discriminative power is still provided by the text information, and particularly by metadata (see T4). The ASR provided by LIMSI [4] tends to be more efficient than the one from LIUM [3] (see T1 vs. T2). Metadata is more reliable than the ASR transcripts (improvement of 14% over ASR; see T3 vs. T2) but cannot be computed automatically (is generated by users, e.g. tags, title). The discriminative power of metadata diminish when mixed with audio-visual information (see AVT5).

2.2 Official runs

We use the SVM linear classifier with: **Run1** - LBP, CCV, hist (visual) and audio block-based [2], **Run2** - TF-IDF on ASR LIMSI [4] (evaluation using only the ASR test set videos), **Run3** - audio block-based [2] + LBP, CCV, hist + TF-IDF on ASR LIMSI [4], **Run4** - audio block-based [2], **Run5** - TF-IDF on metadata + ASR LIMSI [4].

Table 1: MAP for official runs.

team	Run1	Run2	Run3	Run4	Run5
ARF	19.41%	21.74%	22.04%	18.92%	37.93%

Evaluated from the perspective of a retrieval system, the best performance is still achieved using the text descriptors (see Run5). The use of metadata provides significant improvement (see Run5 vs. Run2). Increasing the number of modalities (audio - Run4, audio-visual - Run1, audio-visual-text - Run3) increases progressively the performance.

3. CONCLUSIONS

We have provided a detailed evaluation of a broad range of content descriptors in the context of video genre classification. Experiments show that automatic descriptors have great potential competing with the human generated ones.

4. REFERENCES

[1] S. Schmiedeke, C. Kofler, I. Ferrané, "Overview of MediaEval 2012 Genre Tagging Task", MediaEval Workshop, Pisa, Italy, 4-5 October 2012.

[2] B. Ionescu, K. Seyerlehner, I. Mironică, C. Vertan, P. Lambert, "An Audio-Visual Approach to Web Video Categorization", Multimedia Tools and Applications, DOI:10.1007/s11042-012-1097-x, 2012.

[3] A. Rousseau, F. Bougares, P. Deléglise, H. Schwenk, Y. Estève, "LIUM's Systems for the IWSLT 2011 Speech Translation Tasks", Int. Workshop on Spoken Language Translation, USA, 8-9 September 2011.

[4] L. Lamel, J.-L. Gauvain, "Speech Processing for Audio Indexing", Int. Conf. on Natural Language Processing, LNCS, 5221, pp. 4-15, Springer Verlag, 2008.