A Visual-based Late-Fusion Framework for Video Genre Classification

Ionuț Mironică[1], Bogdan Ionescu[1,2], Christoph Rasche[1], Patrick Lambert[2]
[1] LAPI, University Politehnica of Bucharest, Romania
{imironica, bionescu, rasche}@alpha.imag.pub.ro
[2] LISTIC, University of Savoie, Annecy, France
patrick.lambert@univ-savoie.fr

***Abstract*— In this paper we investigate the performance of visual features in the context of video genre classification. We propose a late-fusion framework that employs color, texture, structural and salient region information. Experimental validation was carried out in the context of the MediaEval 2012 Genre Tagging Task using a large data set of more than 2,000 hours of footage and 26 video genres. Results show that the proposed approach significantly improves genre classification performance outperforming other existing approaches. Furthermore, we prove that our approach can help improving the performance of the more efficient text-based approaches.**

I. INTRODUCTION

The main challenge of the existing multimedia systems is in their capability of identifying and selecting only relevant information according to some user specifications. This issue became more critical because of the tremendous increase of multimedia content. Thanks to new technologies in portable multimedia terminals, wireless transmission protocols and imaging devices, we currently face a large amount of content that is infeasible for humans to manually analyze in view of finding specific information. To have an idea of this scale, the following numbers are edificatory: over 72 hours of video are uploaded on YouTube (http://www.youtube.com/) every minute, more than 500 years of YouTube videos are watched every day on Facebook (https://www.facebook.com/) and over 700 YouTube videos are shared on Twitter (https://twitter.com/) each minute. One of the strategies adopted to narrow the choices is to provide preliminary automatic categorization according to some pre-defined categories, such as *video genres*.

A common approach used in video categorization is the use of *text-based information*. Basically, all existing web media search engines (e.g., YouTube, blip.tv) rely on the use of synopsis, user tags, metadata, etc., to perform the classification. But not all these descriptors can be automatically extracted and they require to be generated manually by the users. In order to obtain the text information automatically, an approach is to extract text from video information, such as movie objects and graphics (e.g., the name of a building) or other visual text (e.g., subtitles). This typically involves the use of Optical Character Recognition (OCR) techniques that might not be that accurate considering the variability of video contents. Another option is to take advantage of the transcripts of dialogues obtained with Automatic Speech Recognition (ASR) techniques. However, many of the movies contain different languages and also background noise that rends ASR highly inefficient.

Another perspective is to use *audio information*. Audio features have the advantage of requiring less computational effort than ASR-based methods. The audio-based information can be derived either from time, e.g., Root Mean Square of signal energy (RMS), sub-band information, Zero-Crossing Rate (ZCR); or from the frequency domain, e.g., bandwidth, pitch, Mel-Frequency Cepstral Coefficients (MFCC).

Visual-based approaches gain more and more popularity in the last years showing great potential in various video retrieval scenarios. They exploit various information sources from color, texture to motion and temporal structure. Giving the complexity of the categorization task, visual features tend to report a lower performance then for text and audio modalities.

In this paper we prove that notwithstanding the superiority of textual and audio descriptors, visual information alone has also great potential to video genre classification tasks. Using the combination of various visual descriptors and adequate fusion mechanisms one can achieve similar performance to the higher level audio and textual features, that is quite a remarkable result.

The remainder of the paper is organized as follows: Section II discuses several relevant genre classification approaches and situates our work accordingly. Section III presents the proposed video descriptors which are extracted from visual information. In Section IV we discuss several fusion techniques and in Section V we report the experimental results. Finally, Section VI concludes the paper.

II. PREVIOUS WORK

In this paper we focus on the visual information as we aim to demonstrate its potential to video genre categorization. A video genre is a set of video documents sharing similar style [1], such as broadcast, home-video or movies.

Many of the existing approaches focused on only few video categories, such as movies, TV programs [2] or online videos [3]. For instance, a simple single modal approach is the one proposed in [4]. It addresses genre classification using only video dynamics. Motion information is extracted at two levels: background camera motion and foreground or object motion. A single feature vector is constituted in the DCT transformed space. This is to assure low-pass filtering, orthogonality and a reduced feature dimension. A Gaussian Mixture Model (GMM)

based classifier is then used to identify 3 common genres: sports, cartoons and news.

Much complex approaches address a broader range of genres such as the method in [3] that exploits several video modalities. At temporal level video contents is described using average shot length, cut percentage, average color difference and camera motion (4 cases are detected: still, pan, zoom, and other movements). Spatial features include face frames ratio, average brightness and color entropy. Genre classification is addressed at different levels, according to a hierarchical ontology of video genres. Several classification schemes (Decision Trees and several SVM approaches) are used to classify video footage into main genres: movie, commercial, news, music and sports; and further into sub-genres, movies into action, comedy, horror and cartoon, and sports into baseball, football, volleyball, tennis, basketball and soccer.

More recently, the Genre Tagging Task of the MediaEval Multimedia Benchmark (http://www.multimediaeval.org/) set up a new perspective for the benchmarking of genre classification approaches. It addressed both large scale categorization and multimodal approaches (text-audio-visual). In 2012 it released a data set that contains up to 26 video genre categories and more than 2,000 hours of footage. Results are discussed in this paper (for a detailed overview see [5]).

In this paper we approach genre categorization using a late-fusion framework of visual descriptors. The validation of the approach is conducted in the context of MediaEval 2012 Genre Tagging Task.

III. VISUAL FEATURES

As previously mentioned, we approach video genre categorization using only visual information. From the existing modalities we exploit *color*, *texture*, *contour* and *salient regions*. Visual content is described by several descriptors:

1) Histogram of Oriented Gradients (HoG, 81 values) [6]: exploits local object appearance and shape within an image via the distribution of edge orientations. This can be achieved by dividing the image into small connected regions (cells) and for each of them building a pixel-wise histogram of edge orientations. For the entire sequence we compute the histogram average of all the frames;

2) MPEG-7 and related descriptors (1,009 values) [7]: we adopted several standard color and texture-based descriptors, namely: Local Binary Pattern (LBP), autocorrelogram, Color Coherence Vector (CCV), Color Layout Pattern (CLP), Edge Histogram (EH), Scalable Color Descriptor (SCD), color histogram and color moments. Overall, sequence aggregation is achieved by computing the mean, dispersion, skewness, kurtosis, median and root mean square statistics of all frames;

3) Structural descriptors (1,430 values) [8]: we propose a new approach that is based on the characterization of geometric attributes of contour information. Contour processing starts with edge detection which is performed with the Canny edge detector. For each contour, a type of curvature space is created. This space is then abstracted into spectra-like functions, from which in turn a number of geometric attributes are derived, such as the degree of curvature, angularity, circularity, symmetry, "wiggliness" and so on. In addition to those geometric parameters, a number of "appearance" parameters are extracted. They consist of simple statistics obtained from the luminance values extracted along the contour, such as the contrast (mean, standard deviation). These descriptors were reported to be successfully employed in tasks such as the annotation of photos and object categorization at ImageCLEF 2010 benchmarking (http://www.imageclef.org/);

4) Bag-of-VisualWords (B-o-VW, 20,480 values): we compute a global B-o-VW model [9] over a selection of frames. The visual vocabulary used was restricted to 4,096 words. Key points are extracted with a dense sampling strategy and described using rgbSIFT features [9]. Descriptors are extracted using the first two levels of a spatial pyramid image representation [10]. This approach reported good performance on image classification and object localization tasks at Pascal Challenge [11].

IV. FUSION APPROACH

To combine the visual descriptors we designed a fusion strategy. In general, there are two types of fusing strategies: *early fusion* and *late fusion*. These strategies are funded around the hypothesis that an aggregated decision from multiple experts can be superior to a decision from a single system.

Early fusion combines features before performing any classification, while late fusion combines the output of classifiers for different descriptors. Each of the considered classifiers, $C_k, k \in \{1,...,K\}$, will have to provide some scores, $x_k = [x_{k1},...,x_{kN}]$, indicating the probability for each video, $v_n, n \in \{1,...,N\}$, to contain the requested concept (e.g, genre in our case). The objective of late fusion is to determine a score combination function, $f()$, so that the resulting aggregated classifier output, $x = f(x_{k1},..,x_{kN})$, is better than any of its individual components, and overall as good as possible. Late fusion focuses on the individual strength of modalities, while early fusion use the correlation of features in the mixed feature space. Numerous empirical and theoretical studies showed that no fusion strategy is optimal in general. The fusion strategy has to be selected according to the task and data structure.

In this paper we propose to use a late fusion strategy because it shows many advantages to our task. First, there is the computational speed, and then its robustness to combine with different types of features or classifiers. In this paper we will evaluate two late fusion techniques, namely: CombSUM and CombMNZ [19]. CombMNZ is determined as:

$$CombMNZ = F(d) \cdot \sum_{n=1}^{N} x_k \alpha_k \quad (2)$$

where *F(d)* is the number of classifiers for which *d* was in the top *K* retrieved videos (its score is considered non zero) and aims to give more importance to the videos that were retrieved by several classifiers, and α_k represent the confidence level of

classifier C_k. If $F(d)=1$ and $\alpha_k=1, k\in\{1,...,K\}$, CombMNZ determines the CombSUM fusion.

V. EXPERIMENTAL RESULTS

A. Experimental Setup

This work focuses on the role of visual information in video genre classification. For experimentation we use the 2012 MediaEval Genre Tagging Task data set [5] that provides genre ground truth for up to 14,838 videos (the data set is divided into a train set of 5,288 videos and a test set of 9,550 movies). Videos are labeled according to 26 video genre categories (for more details see [5]). The main challenge of this task is in the high diversity of genres as well as in the high variety of visual contents within each genre category. Moreover, genre related information tends to be contained mainly with the audio information (e.g., speech) which makes the task of using visual information very challenging.

For genre classification we have selected three of the most popular approaches that proved very efficient in various information retrieval tasks [1], namely: Support Vector Machines (SVM; with various kernel functions), k-Nearest Neighbor (k-NN) and Random Trees (RT).

To assess performance, we use as global measure the classic Mean Average Precision (MAP) that provides a single-figure measure of quality across recall levels:

$$AP=\frac{1}{M}\sum_{n=1}^{N}\frac{f_c(v_n)}{n} \qquad (2)$$

where N is the number of videos, M is the number of videos of genre c, and v_i is the i-th video in the ranked list $\{v_1,...,v_N\}$ returned by the classifier; finally, $f_c()$ is a function which returns the number of videos of genre c in the first i videos if v_i is of class c, and 0 otherwise.

B. Visual feature performance

In *Table I* we present the classification results obtained with each individual visual descriptor. One can observe that using global HoG alone in combination with SVM and a nonlinear kernel we obtain the highest MAP of 26.5%. Similar performance is obtained with MPEG-7 features using RT.

TABLE I. VISUAL FEATURES PERFORMANCES (MAP VALUES).

Feature	***Classification algorithm***				
	SVM Linear	SVM RBF	SVM Chi	k-NN (k=5)	RT
Bag of Words	17.10	19.56	***19.99***	16.22	18.15
Global HOG	9.07	***26.5***	22.24	18.37	23.46
Structural	7.55	17.16	***22.75***	8.75	15.67
MPEG-7 related	6.11	4.45	17.49	9.57	***25.68***
All combined	**18.21**	**30.11**	**31.21**	15.12	21.22

Combining all the features using early fusion, the performance is superior in most of the cases. It proves that each category of visual feature contains complementary information that provides additional discriminative power to the aggregated descriptor.

Fusion techniques exploit the diversity between different information sources. *Table II* reports the results obtained with the proposed late fusion framework, namely: CombSUM and CombMNZ [19]. In both cases, late fusion provides better results then the use of early fusion, with more then 6%. The late fusion is less computational expensive then early fusion because of the descriptor length which is significantly shorter. Furthermore, late fusion scales up easier because no re-training is necessary if further streams or modalities are to be integrated.

The best performance is obtained with late fusion CombMNZ using a combination of top 4 best pairs of feature/classifier from *Table I* (marked with bold italic).

TABLE II. LATE FUSION PERFORMANCE (MAP VALUES).

Method	***MAP***
Late Fusion CombMNZ	**38.33%**
Late Fusion CombSUM	36.89%

C. Comparison to MediaEval 2012 audio-visual approaches

A summary of the best team audio-visual runs at MediaEval 2012 Video Genre Tagging Task is presented in *Table III* (results are presented by decreasing MAP values). The KIT team obtained the best results for visual descriptors. They obtained a MAP of 30.08% using a combination of classical color and texture descriptors: HSV color histogram (162 bins), color moments (Lab color space), autocorrelogram, co-occurrence texture, wavelet texture grid and edge histogram. By combining these features with B-o-VW of rgbSIFT descriptors, they increase the performance to 34.99%. Results show that using only B-o-VW, in spite of its reported high performance in many retrieval tasks, results are not that accurate (MAP is only 23.29% using SIFT and 23.01% with SURF-PCA).

TABLE III. MEDIAEVAL 2012 BEST TEAM RESULTS USING AUDIO-VISUAL INFORMATION (MAP VALUES).

Team	***Descriptor and method***	***MAP***
KIT [17]	visual descriptors (color, texture, rgbSIFT) with SVM	34.99%
KIT [17]	visual descriptors (color, texture) with SVM	30.08%
KIT [17]	visual descriptors (rgbSIFT) with SVM	23.29%
TUB [14]	clustered SURF-PCA with SVM (histogram intersection kernel)	23.01%
ARF [15]	visual (LBP, CCV, color HSV histogram) and audio block-based with Linear SVM	19.41%
UNICAMP UFMG [16]	visual (HMP) and audio with KNN	12.38%
proposed	**CombMNZ with HOG, Structural, MPEG7 related features**	**38.29%**

The proposed approach provides the highest performance in all the cases and we achieve an increase of performance of more than 3% compared to the best approach (see *Table III*).

D. Comparison to MediaEval 2012 text-video approaches

In this section we compare our approach to the best team runs that employed truly multimodal approaches including also

textual information. Results are presented in *Table IV*. The most efficient approach remains the inclusion of metadata information (e.g., video title, tags and user descriptions) as it provides a higher semantic level of description than audio-visual information. This yields the highest MAP of 52.55%. The second and the third best ranked approaches have also used a combination of ASR and metadata, obtaining a performance of 37.93% and 36.75%, respectively. The best reported visual approach come in the fourth position with a MAP of 34.99%.

Although used independently visual features cannot outperform the inclusion of textual data, visual information contains complementary information that can be used to improve the overall performance. To support our hypothesis, we have conducted the following experiment. We describe metadata content using a standard Term Frequency-Inverse Document Frequency (TF-IDF) approach [20]. Then, we combine visual and metadata-based TF-IDF representation. This leads to an increase of performance of 8%. Regardless the highly representative power of textual data, results show that performance can still be improved by including visual information.

TABLE IV. MEDIAEVAL 2012 BEST TEAM RESULTS USING THE INTEGRATION OF TEXT INFORMATION (MAP VALUES).

Team	***Descriptor and method***	***MAP***
TUB[11]	BoW metadata & ASR with SVM	**52.55%**
ARF[12]	TF-IDF metadata & ASR	37.93%
TUD-MM[15]	TF on Visual words, ASR & metadata	36.75%
KIT[14]	visuals descriptors (color, texture, rgbSIFT)	34.99%
TUD[13]	TF-IDF ASR	25.00%
UNICAMP-UFMG[16]	visual B-o-VW	21.12%
proposed	**CombMNZ with HOG, Structural and MPEG7 related features**	**38.29%**
proposed	**CombMNZ with visual and metadata features**	**60.27%**

VI. CONCLUSIONS

In this paper we addressed the role of visual information and fusion mechanisms to video genre classification. We evaluate our framework on a large video data set, namely the MediaEval 2012 Genre Tagging Task data set consisting of 26 genres and more then 2,000 hours of footage.

The proposed visual-based approach outperformed all the visual-based approaches reported at MediaEval 2012. We also report similar results to the more representative text and audio features and we demonstrate that visual information can be used to improve the performance of high level textual features.

In addition, we compared various early and late fusion schemes for combining different descriptors. Results showed that to this task, late fusion is more appropriate due to its reduced computational complexity and robustness. In future work we will adapt the method to address a higher diversity of video categories and extend the framework to other modalities, such as using spatio-temporal information.

ACKNOWLEDGMENTS

Part of this work was carried out by Ionuț Mironică during his research stage at the MHUG group at University of Trento, Italy. We would like to thank Prof. Nicu Sebe and Dr. Jasper Uijlings for the very constructive discussions. Also, Bogdan Ionescu's contribution was supported by the research grant EXCEL POSDRU/89/1.5/S/62557. Finally, we acknowledge the 2012 Genre Tagging Task of the MediaEval Multimedia Benchmark for the data set (http://www.multimediaeval.org/).

REFERENCES

[1] B. Ionescu, K. Seyerlehner, I. Mironica, C. Vertan, P. Lambert, "An Audio-Visual Perspective on Automatic Web Media Categorization", Media Tools and Applications, 2012

[2] D. Borth, J. Hees, M. Koch, A. Ulges, C. Schulze: "Tubefiler: "An automatic web video categorizer". In ACM Multimedia, 2009.

[3] X. Yuan, W. Lai, T. Mei, X. sheng Hua, X. Qing Wu, S. Li "Automatic video genre categorization using hierarchical SVM" in ICIP, 2006.

[4] M.J. Roach, J.S.D. Mason, "Video Genre Classification using Dynamics," IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 1557-1560, Utah, USA, 2001.

[5] S. Schmiedeke, C. Kofler, I. Ferrané, "Overview of MediaEval 2012 Genre Tagging Task", Working Notes of the MediaEval 2012

[6] O. Ludwig, D. Delgado, V. Goncalves, U. Nunes: "Trainable Classifier-Fusion Schemes: An Application To Pedestrian Detection", IEEE Int. Conference On Intelligent Transportation Systems, pp. 432-437, 2009.

[7] T. Sikora: "The MPEG-7 Visual Standard for Content Description - An Overview", IEEE Transactions on Circuits and Systems for Video Technology, pp. 696 - 702, 2001.

[8] C. Rasche: "An Approach to the Parameterization of Structure for Fast Categorization", Int. Journal of Computer Vision, 87, pp. 337-356, 2010.

[9] J.R.R. Uijlings, A.W.M. Smeulders, R.J.H. Scha: "Real-Time Visual Concept Classification", IEEE Transactions on Multimedia, 99, 2010

[10] S. Lazebnik, C. Schmid, J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories", in IEEE Conference on Computer Vision and Pattern Recognition, 2006

[11] M. Marszałek, C. Schmid, H. Harzallah, and J. van de Weijer, "Learning representations for visual object class recognition," ICCV, 2007

[12] C. G. M. Snoek, M. Worring, A. W. M. Smeulders "Early versus late fusion in semantic video analysis", ACM Multimedia, 2005

[13] A. F. Smeaton, P. Over, W. Kraaij: "High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements", on Multimedia Content Analysis Theory and Applications, 2009.

[14] S. Schmiedeke, P. Kelm, T. Sikora, "TUB @ MediaEval 2012 Tagging Task: Feature Selection Methods for Bag-of-(visual)-Words Approaches", Working Notes Proc. of the MediaEval 2012 Workshop.

[15] B. Ionescu, I. Mironica, K. Seyerlehner, P. Knees, J. Schluter, M. Schedl, H. Cucu, A. Buzo, P. Lambert "ARF @ MediaEval 2012: Multimodal Video Classification", Working Notes of the MediaEval 2012 Workshop.

[16] Y. Shi, M. A. Larson and C. M. Jonker, "MediaEval 2012 Tagging Task: Prediction based on One Best List and Confusion Networks", Working Notes Proc. of the MediaEval 2012 Workshop.

[17] T. Semela, M. Tapaswi, H. K.l Ekenel, R, Stiefelhagen, "KIT at MediaEval 2012 - Content-based Genre Classification with Visual Cues", Working Notes Proc. of the MediaEval 2012 Workshop.

[18] Peng Xu, Yangyang Shi and Martha Larson, "TUD at MediaEval 2012 genre tagging task: Multi-modality video categorization with one-vs-all classifiers", Working Notes Proc. of the MediaEval 2012 Workshop.

[19] G. Csurka, S. Clinchant, "An empirical study of fusion operators for multimodal image retrieval", International Workshop on Content-Based Multimedia Indexing, Annecy, France, 2012.

[20] P. Knees, T. Pohle, M. Schedl, D. Schnitzer, K. Seyerlehner, G. Widmer, "Augmenting Text-Based Music Retrieval with Audio Similarity", International Society for Music Information Retrieval, ISMIR 2009.