An In-Depth Evaluation of Multimodal Video Genre Categorization

Ionuţ Mironică[1], Bogdan Ionescu[1,3], Peter Knees[2], Patrick Lambert[3]
[1] LAPI, University ”Politehnica” of Bucharest, 061071, Romania
Email: {imironica, bionescu}@alpha.imag.pub.ro
[2] Johannes Kepler University, 4040 Linz, Austria
Email: peter.knees@jku.at
[3] LISTIC, Polytech Annecy-Chambery, University of Savoie, 74944, France
Email: patrick.lambert@univ-savoie.fr

***Abstract*—In this paper we propose an in-depth evaluation of the performance of video descriptors to multimodal video genre categorization. We discuss the perspective of designing appropriate late fusion techniques that would enable to attain very high categorization accuracy, close to the one achieved with user-based text information. Evaluation is carried out in the context of the 2012 Video Genre Tagging Task of the MediaEval Benchmarking Initiative for Multimedia Evaluation, using a data set of up to 15.000 videos (3,200 hours of footage) and 26 video genre categories specific to web media. Results show that the proposed approach significantly improves genre categorization performance, outperforming other existing approaches. The main contribution of this paper is in the experimental part, several valuable interesting findings are reported that motivate further research on video genre classification.**

I. Introduction

Searching for video content proves to be in many cases a tendinous task. This is mainly due to the lack of appropriate tools for handling, on one side the rich multimodal video contents, and on the other side the constantly growing large collections of video footage. To bridge this gap, existing media platforms (e.g., YouTube, Dailymotion, blip.tv) are still relying on user generated video categorization (e.g., based on user tags or metadata). The design of a system that would handle this task automatically is still an open challenge.

Machine learning techniques have been used extensively to solve various video categorization tasks as they can handle massive data, e.g., high descriptor dimensionality, tens of thousands of items. However, most of the approaches reported in the literature are limited to address very few categories, such as determining classic TV genres, e.g., action, comedy, horror, cartoon, sports. Best performance is reported with multimodal approaches that exploit the benefits of fusing various modalities: text, visual and audio.

From the modality point of view, the use of user generated textual information (e.g., synopsis, user tags, metadata) provides the most accurate results. The main drawback is that it cannot be generated automatically, which limits its applications. Text can also be obtained in an automated manner, either from scene text (e.g., graphic text, sub-titles), or from the transcripts of dialogs obtained with Automatic Speech Recognition (ASR) techniques [27]. Video footage however may contain different languages and also background noise that rends ASR highly inefficient. Video categorization using text is typically accomplished with classic Bag-of-Words model and Term Frequency-Inverse Document Frequency (TF-IDF) approaches.

Although reported as being less accurate than text, the use of audio-visual information is the most popular choice, mainly because it can be derived directly from the video footage itself. Audio-based information can be extracted from, both, time and frequency domains. Common time-domain approaches include the use of Root Mean Square of signal energy, sub-band information, Zero-Crossing Rate or silence ratio; while frequency-domain features include energy distribution, frequency centroid, bandwidth, pitch or Mel-Frequency Cepstral Coefficients - MFCC (see Yaafe audio features extraction toolbox http://yaafe.sourceforge.net/).

Visual-based information exploits both static and dynamic aspects either in the spatial domain using color, temporal structure, objects, feature points, motion, or in the compressed domain, e.g., using MPEG coefficients. Some of the most efficient approaches use feature points, e.g., Scale Invariant Feature Transform (SIFT) [28], Space-Time Interest Points (STIP) [1], Histogram of oriented Gradients (HoG), 3D-SIFT [2], and Bag-of-Visual-Words representations [3]. These methods are however known to be very computational expensive due to the computation of the visual word dictionaries.

In this paper we propose an in-depth evaluation of the performance of video descriptors to multimodal video genre categorization. We discuss the perspective of designing appropriate late fusion techniques that would enable to attain very high categorization accuracy, close to the one achieved with user-based text information.

The remainder of the paper is organized as follows. Section II discuses several relevant video genre categorization approaches and situates our work accordingly. The proposed multimodal descriptors and fusion strategies are presented in Section III and Section IV, respectively. Section V reports the experimental results. Finally, Section VI provides a brief summary and concludes the paper.

II. Previous work

Automatic video genre categorization has been studied extensively in the literature from now more than ten years

[5]. Most of the work focuses on the categorization of movie genres, TV broadcasting [6] or online videos [7]. Existing approaches range from exploiting single-modality to multimodal integration.

For instance, the approach in [8] uses only text-based information. It proposes an incremental Support Vector Machine (SVM) approach that makes use of online Wikipedia propagation to categorize large-scale web videos. It combines contextual and social information, such as metadata, user behavior, viewer's behavior and video relevance. A visual-based approach is the one in [9] that proposes a framework for distinguishing from two different level of concepts: a genre-specific concept classification layer and a frame-concept detection module. For genre classification, visual content is described using Bag-of-Visual-Words representation of Opponent SIFTs that are classified with a probabilistic model. [6] proposes the use of text and visual information for web video categorization. A genre-based categorization is first achieved using video tags and title, while sub-genres are further determined using visual features. Genre classification is addressed at different levels, according to a hierarchical ontology of genres.

A multimodal approach that considers also audio information is proposed in [5]. It combines synchronized audio MFCCs features and mean and standard deviation of motion vectors and MPEG-7 visual descriptors. Videos are classified with a Gaussian Mixture Models (GMM) based classifier. Another example is the approach in [10] that extracts features from four information sources: visual-perceptual information (color, texture, and motion), structural information (shot length, shot distribution, shot rhythm, shot clusters duration, and saturation), cognitive information (e.g., numbers, positions, and dimensions of faces), and aural information (transcribed text, sound characteristics). These features are used to train a parallel Neural Network system and used to distinguish between seven TV genres. An interesting categorization approach is the cross-media retrieval proposed in [11]. It is founded on the construction of a Multimedia Correlation Space (MMCS) which exploits semantic correlations between different modalities based on content description and co-occurrence information. The proposed video descriptors are related to color, texture and aural information.

Systems that combine multiple features using various fusion techniques have been proved to improve performance in various classification tasks [29] [30]. However, for video genre categorization most of the existing approaches are limited to make use of simple descriptor concatenation.

More recently the Genre Tagging Task of the MediaEval Benchmarking Initiative for Multimedia Evaluation [12] set up a new perspective for benchmarking video genre categorization by addressing, both, large scale categorization (data set of up to 15,000 Internet videos) and multimodal approaches (text-audio-visual). It targets a real-world scenario, i.e., the video genre categorization employed by the blip.tv video platform.

In this paper we present an in-depth evaluation study of multimodal video genre categorization. We investigate the use of various state-of-the-art content descriptors extracted from different modalities and the efficiency of early and late fusion mechanisms to this task. With this work we attempt to respond to several research questions, such as to what extent aural and visual information can lead to similar performance or even surpass the highly semantic textual descriptors? How efficient would be an adequate combination of various modalities in achieving highly accurate classification? How really important is the contribution of video modalities in improving the accuracy of textual data. Experimentation is carried out in the context of the 2012 Genre Tagging Task of the MediaEval Benchmarking [12].

Although the proposed methods have been more or less previously explored in the literature, the main contributions of this work are the following: (1) we provide an in-depth evaluation of truly multimodal video description (automated and user-generated text, audio and visual) in the context of a real-world genre-categorization scenario; (2) we demonstrate the potential of appropriate late fusion to genre categorization and achieve very high categorization performance; (3) we prove that notwithstanding the superiority of user-text based descriptors, late fusion can boost performance of automated content descriptors to achieve close performance; (4) we setup a new baseline for the Genre Tagging Task by outperforming the performance of the other participants; (5) evaluation is carried out on a standard data set [12] making the results both relevant and reproducible.

III. Content description

It is well know that different modalities tend to account for different information providing complementary discriminative power. We experiment all the available sources of information, from the audio, visual, to highly semantic textual information obtained with Automatic Speech Recognition (ASR) as well as user generated data (e.g., metadata that typically accompanies video content on the Internet).

Aural information. Common video genres have very specific audio signatures, e.g., documentaries use a mixture of natural sounds and monologues, music clips contain different music genres, sports feature crowd noise and also monologues, in talk shows speech is predominant. To describe these aspects, we employ a set of standard audio features that provided good performance in audio genre categorization tasks [5]:

-*standard audio features (196 values)* [22] contains a common set of general-purpose audio descriptors, namely: Linear Predictive Coefficients, Line Spectral Pairs, MFCCs, Zero-Crossing Rate, 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 taking 1.28 s). Video temporal integration is achieved by taking the mean and standard deviation of these descriptors over all frames.

Visual information. From the visual point of view, the distribution of color and feature/object information highlight specific genre contents. For instance, music clips and movies tend to have darker colors, commercials use many visual effects, sports usually have a predominant hue, newscasts show a high frequency of faces, documentaries, sports and news have specific contour signatures, e.g., skyline contours, people silhouettes. To capture those particularities, we determine the following state-of-the-art descriptors that are classic approaches successfully employed in various image/video retrieval scenarios [13] [14] [26]:

-MPEG-7 related descriptors (1,009 values) [23] describe the global color and texture information over all the frames. We selected the following representative descriptors: Local Binary Pattern, autocorrelogram, Color Coherence Vector, Color Layout Pattern, Edge Histogram, Scalable Color Descriptor, classic color histogram 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.

-structural descriptors (1,430 values) account for contour information and relation between contours. We use the approach in [25] to parameterize the geometry and appearance of contours and regions of gray-scale images with the goal of object categorization. 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, e.g., the degree of curvature, angularity, circularity, symmetry, "wiggliness".

-global Histograms of oriented Gradients (HoG - 81 values) represent the average of the well known HoG features [24]. It 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. The combination of these histograms represent the descriptor which has the advantage of providing contextual information. For the entire sequence we compute the average histogram over all frames.

-Bag-of-VisualWords of SIFT descriptors (20,480 values), we compute a Bag-of-VisualWords (B-o-VW) model over a selection of key frames (uniformly sampled). For this task, we extract a visual vocabulary of 4,096 words. The keypoints are extracted with a dense sampling strategy and described using rgbSIFT features [28]. Descriptors are extracted at two different spatial scales of a spatial pyramidal image representation (entire image and quadrants).

Textual information. Textual data is by far the most representative for providing content information. Specific keywords, e.g., "religion", "economy", "music", can reveal meaningful information about genres. For instance, metadata usually contains the video title, user tags, comments and content descriptions that are highly correlated to genre concepts. For text description we adapted a classic Term Frequency-Inverse Document Frequency (TF-IDF) approach. First, we filter the input text by removing the 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 with the highest χ^2 values that occur more frequently than in complement classes. This results in a vector representation for each video sequence that is subsequently cosine normalized to remove the influence of the length of text data. We consider following TF-IDF descriptors:

-TF-IDF of ASR data (3,466 values, $m = 150$) describes textual data obtained from Automatic Speech Recognition of the audio signal. For ASR we use the transcripts provided by [27] that proved highly efficient to genre classification [12].

-TF-IDF of metadata (504 values, $m = 20$) describes textual data obtained from user metadata such as synopsis, user tags, video title, information that typically accompanies videos posted on the blip.tv platform.

IV. Multimodal integration

To combine various information sources one has to deploy efficient fusion strategies. In general, there are two approaches: *early fusion* and *late fusion* [29]. These strategies are based on the hypothesis that an aggregated decision from multiple classifiers can be superior to a decision from a single one.

The early fusion combines the descriptors before the classification. The combination takes place in the feature space, namely the features are concatenated into one vector. The major drawback in this case is the dimensionality of the resulting feature space that is basically the sum of all the concatenated dimensions. High-dimensional spaces tend to scatter the homogeneous clusters of instances belonging to the same concepts reducing the performance.

In contrast, late fusion combines the confidence values from classifiers run on different descriptors. In our scenario, a classifier is supposed to provide some relevance scores indicating the probability of each sequence of belonging to some class (i.e., genre in our case). Naturally, each of the classifiers will tend to provide different scores for each class. Late fusion involves the design of an aggregated classifier combination function, $f(x_1, ..., x_N)$, with x_i the relevance output of the classifier i, whose result is better than any of its individual classifiers and as good as possible. The aggregation is carried out for each individual class. To achieve the final categorization, videos are then sorted according to the new aggregated scoring formula results.

Late fusion focuses on the individual strength of modalities, whereas early fusion use the correlation of features in the mixed feature space. There are many strategies for late fusion, such as weighted majority voting, rank-level fusion methods or fusion techniques that uses second tier classifiers [29] [30] [13]. However, there is no reported supremacy of one approach over the other, results show that typically the fusion mechanism is adapted to the specificity of each task [31].

For our study we have selected several popular approaches that are presented in the following. For each video document and feature pair, classification yields C confidence scores, one for each of the target genres. Simple linear combination represents a weighted average of the multimodal confidence values of each of the considered classifiers and is defined as:

$$CombMean(d,q) = \sum_{i=1}^{N} \alpha_i \cdot cv_i \tag{1}$$

where cv_i is the confidence value of classifier i for class q ($q \in \{1, ..., C\}$), d is the current video, α_i are some weights and N is the number of classifiers to be aggregated. In case of considering equal weights, $\alpha_1 =, ..., = \alpha_N$, this is referred to as CombSum.

An extension of CombMean can be obtained by giving more importance to the documents that are more likely to be

Table I. CLASSIFICATION PERFORMANCE OF INDIVIDUAL MODALITIES (MAP).

Descriptors	**SVM Linear**	**SVM RBF**	**SVM CHI**	**5-NN**	**RF**	**ERF**
HoG	9.08 %	**25.63%**	22.44%	17.92%	16.62%	23.44%
Bag-of-Visual-Words rgbSIFT	14.63 %	17.61%	**19.96%**	8.55%	14.89%	16.32%
MPEG-7	6.12 %	4.26%	17.49%	9.61%	20.90%	**26.17%**
Structural descriptors	7.55 %	17.17%	**22.76%**	8.65%	13.85%	14.85%
Standard audio descriptors	20.68 %	24.52%	35.56%	18.31%	34.41%	**42.33%**
TF-IDF of ASR	32.96 %	**35.05%**	28.85%	12.96%	30.56%	27.93%
TF-IDF of metadata	56.33 %	58.14%	47.95%	57.19%	**58.66%**	57.52%

relevant for current concepts, which leads to:

$$CombMNZ(d,q) = F(d)^{\gamma} \cdot \sum_{i=1}^{N} \alpha_i \cdot cv_i \tag{2}$$

where $F(d)$ is the number of classifiers for which video d appears in the top K of the retrieved videos and $\gamma \in [0,1]$ is a parameter.

Finally, another useful perspective is to consider the rank of each confidence level. The score-based late fusion strategies require a normalization among all confidence values in order to balance the importance of each of them, which is not the case of the rank-based strategies. In our scenario, we use a common method for rank-based fusion, that is Borda Count. The document with the highest rank on each rank-list gets n votes, where n is the size of the dataset:

$$CombRank(d,q) = \sum_{i=1}^{N} \alpha_i \cdot rank(cv_i) \tag{3}$$

where $rank()$ represents the rank of classifier i, α_i are some weights and N is the number of classifiers to be aggregated.

V. EXPERIMENTAL RESULTS

Experimentation was conducted in the context of the MediaEval Benchmarking Initiative for Multimedia Evaluation, 2012 Genre Tagging Task [12].

The data set consisted of up to 14,838 blip.tv videos that were divided into a training set of 5,288 videos (36%) and a test set of 9,550 movies (64%; we use the same scenario as for the official benchmark). Videos are labeled according to 26 video genre categories specific to the blip.tv media platform, namely (the numbers in brackets are the total number of videos): art (530), autos and vehicles (21), business (281), citizen journalism (401), comedy (515), conferences and other events (247), documentary (353), educational (957), food and drink (261), gaming (401), health (268), literature (222), movies and television (868), music and entertainment (1148), personal or auto-biographical (165), politics (1107), religion (868), school and education (171), sports (672), technology (1343), environment (188), mainstream media (324), travel (175), video blogging (887), web development an (116) and default category (2349, comprises movies that cannot be assigned to any of the previous categories). The main challenge of this scenario is in the high diversity of genres, as well as in the high variety of visual contents within each genre category (for more details see [12] [15]).

For classification, we have selected five of the most popular approaches that proved to provide high performance in various information retrieval tasks [12] [13] [15] [26] [14], namely Support Vector Machines (SVM, with various kernel functions: linear, Chi-square - CHI, Radial Basis Functions - RBF), k-Nearest Neighbor (k-NN), Random Trees (RT) and Extremely Random Forest (ERF).

To assess performance, we report the standard Mean Average Precision (MAP) that is computed as the average value of the Average Precision:

$$AP = \frac{1}{m} \sum_{k=1}^{n} \frac{f_c(v_k)}{k} \tag{4}$$

where n is the number of videos, m is the number of videos of genre c, and v_k is the k-th video in the ranked list $\{v_1, ..., v_n\}$. Finally, $f_c()$ is a function which returns the number of videos of genre c in the first k videos if v_k is of genre c and 0 otherwise (we used the trec_eval scoring tool available at http://trec.nist.gov/trec_eval/).

A. Performance assessment of individual modalities

The first experiment consisted on assessing the discriminative power of each individual modality and group of descriptors. Table I summarizes some of the results (the best performance per modality is highlighted in bold).

The highest performance for visual information is achieved using MPEG-7 related descriptors and Extremely Random Forest (ERF) classifiers, MAP 26.17%, followed closely by HoG histograms on SVM and RBF kernel, MAP 25.63%. Surprisingly, Bag-of-Visual-Words representation of feature information (rgbSIFT) is not performing efficiently to this task, MAP is below 20%. The audio descriptors are able to provide a significantly higher discriminative power, the highest MAP of 42.33% being achieved with the ERF classifier.

In what concerns the text modality, the use of metadata and Random Forest classifiers led to the highest MAP of 58.66% which is an improvement of more than 16% over the audio. The use of ASR data alone is able only to provide a MAP up to 35.05% (with SVM and RBF kernel), which is less discriminative than using audio descriptors. Therefore, video descriptors can outperform at this point the automated text descriptors. This is mostly due to the fact that ASR data is extracted automatically, being inherently subject to errors (e.g., due to background noise).

From the classifier point of view, regardless the modality, the lowest performance tends to be obtained with SVM Linear and 5-NN classifiers. This proves that these video features have restraint linear separability and solving the genre classification problem will require more complex nonlinear classification schemes. In the following we investigate the advantage of combining different modalities as well as the impact of the fusion scheme.

Table II. COMPARISON WITH MEDIAEVAL 2012 VIDEO GENRE TAGGING TASK BEST RUNS [12] (MAP)

Team	Modality	Method	MAP
proposed	all	Late Fusion CombMNZ with all descriptors	**65.82**%
proposed	text	Late Fusion CombMean with TF-IDF of ASR and metadata	**62.81**%
TUB [16]	text	Naive Bayes with Bag of Words on text (metadata)	52.25%
proposed	all	Late Fusion CombMNZ with all descriptors except for metadata	**51.9**%
proposed	audio	Late Fusion CombMean with standard audio descriptors	**44.50**%
proposed	visual	Late Fusion CombMean with MPEG-7 related, structural, HoG and B-o-VW with rgbSIFT	**38.21**%
ARF [17]	text	SVM linear on early fusion of TF-IDF of ASR and metadata	37.93%
TUD [20]	visual & text	Late Fusion of SVM with B-o-W (visual word, ASR & metadata)	36.75%
KIT [19]	visual	SVM with Visual descriptors (color, texture, B-o-VW with rgbSIFT)	35.81%
TUD-MM [18]	text	Dynamic Bayesian networks on text (ASR & metadata)	25.00%
UNICAMP - UFMG [21]	visual	Late fusion (KNN, Naive Bayes, SVM, Random Forests) with BOW (text ASR)	21.12%
ARF [17]	audio	SVM linear with block-based audio features	18.92%

B. Performance of multimodal integration

Fusion techniques tend to exploit complementarity among different information sources. In this experiment we assess the performance of various combination of modalities as well as of different fusion strategies, from late fusion schemes (see Section IV) to the simple concatenation of different descriptors (i.e., early fusion).

For late fusion, weights (i.e., α_k and $F(d)$ values) are first estimated on the training set and tuned for best performance. To avoid overfitting, half of the training set is used for training and the other half for parameter evaluation. The actual classification is then carried out on the test set. MAP is reported in Table III (highest values per feature type are presented in bold).

Table III. PERFORMANCE OF MULTIMODAL INTEGRATION (MAP).

Descriptors	Comb SUM	Comb Mean	Comb MNZ	Comb Rank	Early Fusion
all visual	35.82%	36.76%	**38.21%**	30.90%	30.11%
all audio	43.86%	44.19%	**44.50%**	41.81%	42.33%
all text	62.62%	**62.81%**	62.69%	50.60%	55.68%
all	64.24%	65.61%	**65.82%**	53.84%	60.12%

In all of the cases, late fusion tends to provide better performance than early fusion. Using only the visual descriptors the improvement is of more then 8% over simple descriptor concatenation (highest MAP is 38.21% using CombSUM). For audio descriptors, highest MAP of 44.5% is achieved with CombMNZ, that is an improvement of more than 2% over the simple use of all descriptors together. Audio still provides significant superior discriminative power than using only visual.

A significant improvement of performance is also achieved for textual descriptors. We obtain the highest MAP score with CombMean, namely 62.81%, which is an improvement of over 7% compared to early fusion. Although the simple concatenation of modalities manages to boost classification performance up to a MAP of 60.12%, late fusion is able to exploit better the complementarity between descriptors, achieving more than 5% of improvement. In what concerns the late fusion techniques, CombRank tends to provide the least accurate results in most of the cases, while the other approaches tend to provide more or less similar results.

Therefore, in most of the cases late fusion proves to be a better choice for multimodal genre classification. Firstly, it provides significantly higher performance than early fusion. Secondly, late fusion is also less computational expensive than early fusion, because the descriptors used for each of the classifiers are shorter than using the concatenation of all features. Finally, late fusion systems scale up easier because no re-training is necessary if further streams or modalities are to be integrated.

C. Comparison to state-of-the-art

The final experiment consisted on comparing the late fusion strategies against other methods from the literature. As reference, we use the best team runs reported at MediaEval 2012 Video Genre Tagging Task [12]. Results are presented in Table II by decreasing MAP values (one should note that the comparison with MediaEval results is indicative as the official runs were developed under time constraints and without a priori knowledge of the test set ground truth).

The most efficient modality remains the exploitation of textual information as it provides a higher semantic level of description than audio-visual information. In particular, the use of metadata proves to be the most efficient approach leading to the highest MAP at MediaEval 2012, 52.25% (see team TUB [16]). In spite of this high classification rate, late fusion still allows for significant improvement, for instance CombMean on ASR and metadata achieves a MAP up to 62.81% - that is an improvement of more than 10% over the best run at MediaEval 2012 and of around 25% over using the same combination of textual descriptors (team ARF [17]).

In what concerns the visual modality, best MAP at MediaEval 2012 is up to 35% (see team KIT [19]) and is obtained using a combination of classical color/texture descriptors (e.g., HSV color histogram, L*a*b* color moments, autocorrelogram, concurrence texture, wavelet texture grid and edge histograms) and B-o-VW of rgbSIFT descriptors. Results show that using only B-o-VW of feature descriptors (e.g., SIFT, SURF - Speeded-up Robust Features), in spite of their reported high performance in many retrieval tasks, is not that accurate, e.g., MAP 23.29% using SIFT, 23.01% with SURF-PCA (see detailed competition results at [12] [16] [19]). The CombMean late fusion of visual descriptors provides an improvement over the best run of more than 3% (MAP 38.21%).

Using only audio information, best reported run at MediaEval 2012 achieves a MAP of 18.92% (see team ARF [17]). In this case CombMean late fusion of audio descriptors provides an improvement of more than 25% (MAP 44.5%).

Combining all the descriptors with CombMNZ we achieve a very high classification accuracy as MAP is up to 65.82%, that is an improvement of more than 13% over the MediaEval 2012 best run. In spite of the high discriminative power of textual descriptors, the combination of all the modalities with late fusion is able to exploit data complementarity at some level as the improvement over using only textual information is of 3%. This is a significant achievement considering the scale of the data set.

From the modality point of view, metadata provides the highest discriminative power for genre categorization. However, one should note that this information is user generated (e.g., includes document title, tags and user comments and descriptions) and cannot be determined automatically from the video information, that limits its applicability in real-time categorization scenarios. Approaching the classification using only content information that can be computed automatically from video data (ASR and audio-visual descriptors), late fusion is still able to provide high classification performance leading to a MAP of 51.9%, surpassing even some metadata-based approaches, e.g., see team ARF [17] and TUD-MM [18].

VI. Conclusions

In this paper we addressed the problem of automatic video genre categorization. We studied the contribution of various modalities and the role of the fusion mechanisms in increasing the accuracy of the results. The study was carried out in a real-world scenario on 26 blip.tv web video categories and more than 3,200 hours of video footage. The design of appropriate descriptors and late fusion integration allows to achieve a MAP up to 65.8%, that is a significant improvement of more than 13% over the best approach reported at the 2012 MediaEval Genre Tagging Task. We prove that notwithstanding the superiority of employing user-generated textual information (e.g., user tags, metadata), the proposed multimodal integration allows to boost performance of automated content descriptors to achieve close performance. Future work will mainly consist in exploring spatio-temporal data representation in this context.

VII. Acknowledgments

Part of this work was carried out by Ionuţ 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] I. Laptev: "On space-time interest points", International Journal of Computer Vision 64 (2), 107-123, 2005.

[2] K. K. Reddy, M. Shah: "Recognizing 50 Human Action Categories of Web Videos", Journal of Machine Vision and Applications, 1-11, 2012

[3] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, C. Bray: "Visual Categorization with Bags of Keypoints", ECCV, 2004.

[4] H. K. Ekenel, T. Semela, and R. Stiefelhagen. "Content-based video genre classification using multiple cues". In International Workshop on Automated Information Extraction in Media Production, pp. 21-26, 2010.

[5] L.-Q. Xu and Y. Li, "Video classification using spatial-temporal features and PCA", ICME, pp. 485-488, 2003.

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

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

[8] Y. Song, Y.-D. Zhang, X. Zhang, J. Cao, J.-T. Li: "Google challenge: Incremental-learning for web video categorization on robust semantic feature space", ACM Multimedia, pp. 1113-1114, 2009.

[9] J. Wu, M. Worring: "Efficient Genre-Specific Semantic Video Indexing", Multimedia, IEEE Transactions on 14 (2), 291-302, 2012.

[10] M. Montagnuolo, A. Messina: "Parallel Neural Networks for Multimodal Video Genre Classification", Multimedia Tools and Applications, pp. 125-159, 2009.

[11] Y. Yang, D. Xu, F. Nie, J. Luo, Y. Zhuang, "Ranking with Local Regression and Global Alignment for Cross Media Retrieval", ACM Multimedia, pp. 175-184, 2009.

[12] S. Schmiedeke, C. Kofler, I. Ferran "Overview of the MediaEval 2012 Tagging Task", Working Notes Proc. of the MediaEval 2012 Workshop, Pisa, Italy, October 4-5, 2012, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-927/mediaeval2012_submission_2.pdf.

[13] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, Alan F. Smeaton, G. Quenot: "TRECVID 2012 - An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics", Proceedings of TRECVID, 2012.

[14] M. Marszaek, C. Schmid, H. Harzallah, J. van de Weijer, "Learning representations for visual object class recognition", ICCV Pascal VOC 2007 challenge workshop, 2007.

[15] B. Ionescu, K. Seyerlehner, I. Mironica, C. Vertan, P. Lambert: "An Audio-Visual Approach to Web Video Categorization", Multimedia Tools and Applications, 1-26, 2012.

[16] 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.

[17] 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 Proc. of the MediaEval 2012 Workshop.

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

[19] 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.

[20] 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.

[21] J. Almeida, T. Salles, E.r Martins, O. Penatti, R. Torres, M. Goncalves, "UNICAMP-UFMG at MediaEval 2012: Genre Tagging Task", Working Notes Proc. of the MediaEval 2012 Workshop.

[22] Yaafe core features, http://yaafe.sourceforge.net/ , last accessed 2013.

[23] 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.

[24] O. Ludwig, D. Delgado, V. Goncalves, U. Nunes: "Trainable Classifier-Fusion Schemes: An Application To Pedestrian Detection", IEEE Int. Conf. On Intelligent Transportation Systems, 1, pp. 432-437, 2009.

[25] C. Rasche: "An Approach to the Parameterization of Structure for Fast Categorization", Int. Journal of Computer Vision, 87(3), pp. 337-356, 2010.

[26] S. Nowak, M. Huiskes: "New strategies for image annotation: Overview of the photo annotation task at ImageClef 2010", In the Working Notes of CLEF 2010.

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

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

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

[30] G. Csurka, S. Clinchant "An empirical study of fusion operators for multimodal image retrieval", In 10th Workshop on Content-Based Multimedia Indexing Annecy, France, 2012.

[31] Z. Lan, L. Bao, S.-I. Yu, W. Liu, A.G. Hauptmann, "Double Fusion for Multimedia Event Detection". In Proc. MMM, Klagenfurt, Austria, 2012.