Content-Based Video Description for Automatic Video Genre Categorization

Bogdan Ionescu[1,3], Klaus Seyerlehner[2], Christoph Rasche[1], Constantin Vertan[1], and Patrick Lambert[3]

[1] LAPI, University "Politehnica" of Bucharest 061071 Bucharest, Romania
{bionescu,rasche,cvertan}@alpha.imag.pub.ro
[2] DCP, Johannes Kepler University, A-4040 Linz, Austria
klaus.seyerlehner@jku.at
[3] LISTIC, University of Savoie, BP 80439, 74944 Annecy-le-Vieux Cedex, France
patrick.lambert@univ-savoie.fr

Abstract. In this paper, we propose an audio-visual approach to video genre categorization. Audio information is extracted at block-level, which has the advantage of capturing local temporal information. At temporal structural level, we asses action contents with respect to human perception. Further, color perception is quantified with statistics of color distribution, elementary hues, color properties and relationship of color. The last category of descriptors determines statistics of contour geometry. An extensive evaluation of this multi-modal approach based on on more than 91 hours of video footage is presented. We obtain average precision and recall ratios within $[87\% - 100\%]$ and $[77\% - 100\%]$, respectively, while average correct classification is up to 97%. Additionally, movies displayed according to feature-based coordinates in a virtual 3D browsing environment tend to regroup with respect to genre, which has potential application with real content-based browsing systems.

Keywords: video genre classification, block-level audio features, action segmentation, color perception, contour geometry, video indexing.

1 Introduction

The automatic labeling of video footage according to genre is a common requirement when dealing with indexing of large and heterogenous collection of video materials. This task may be addressed, either *globally*, or *locally*. Global-level approaches aim at classifying videos into one of several main genres, e.g. cartoons, music, news, sports, documentaries, etc.; or even more fine-grained into sub-genres, e.g. identifying specific types of sports (football, hockey, etc.), movies (drama, thriller, etc.), and so on. On the other hand, with local-level approaches video segments are labeled according to specific human-like concepts, e.g. outdoor vs. indoor scenes, action segments, violence scenes, etc. (see TRECVid campaign [1]). In this paper we focus on the global classification task only and video genre classification is consequently interpreted as a typical machine learning problem that involves two fundamental steps: *feature extraction* and *data*

K. Schoeffmann et al. (Eds.): MMM 2012, LNCS 7131, pp. 51–62, 2012.

classification. Especially the choice of a suitable task specific feature set is critical for the success of such a classification approach and an ideal feature set should contain as many genre specific cues as possible. In the literature so far, various sources of information have been exploited [2]. Some sources of information may provide more informative cues than others, like for instance visual elements compared to text or even some audio descriptors. The most reliable approaches (which also target the wider range of genres) are however *multi-modal*, i.e. multi-source.

In the following we shall highlight the performance of several approaches we consider relevant for the present work. A simple, single modal approach is the one proposed in [3]. It addresses the genre classification task 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. Despite the limited content information used, when applied to a reduced number of genres, it is able to achieve detection errors below 6%.

A much more complex approach which uses spatio-temporal information is proposed in [4]. 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. The genre classification task 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 into sub-genres, movies into action, comedy, horror and cartoon, and finally sports into baseball, football, volleyball, tennis, basketball and soccer. The highest precision for video footage categorization is around 88.6%, while for sub-genres, sports categorization achieve 97% and movies up to 81.3%.

A truly multi-modal approach, which combines several categories of content descriptors, is proposed in [5]. Features are extracted from four informative sources, which include visual-perceptual information (color, texture and motion), structural information (shot length, shot distribution, shot rhythm, shot clusters duration and saturation), cognitive information (face properties, such as number, positions and dimensions) and aural information (transcribed text, sound characteristics). These features are used for training a parallel Neural Network system and achieve an accuracy rate up to 95% in distinguish between seven video genres, namely: football, cartoons, music, weather forecast, newscast, talk shows and commercials.

In our approach, we exploit for genre classification both audio and visual modalities. The proposed set of *audio features* are block-level based, which compared to classic approaches, e.g. Mel-Frequency Cepstral Coefficients - MFCC [6], have the advantage of capturing local temporal information by analyzing sequences of consecutive frames in a time-frequency representation. On the other

hand, *visual information* is described with temporal information, color and contour geometry. Temporal descriptors are first derived using a classic confirmed approach, i.e. analyzing the frequency of shot changes [4]. However, the novelty is in the way we measure action content, which is based on the assessment of action perception. Color information is extracted globally. Compared to most of the existing approaches which use mainly local or low-level descriptors, e.g. predominant color, color variance, color entropy, frame based histograms [2], the novelty of our approach is in the analysis of color perception. Using a color naming system, color perception is quantified in terms of statistics of color distribution, elementary hues distribution, color properties (e.g. amount of light colors, cold colors, saturated colors, etc.) and relationship of color. The final visual descriptors are related to contour information, which was hardly exploited with genre classification [2]. Instead of describing closed region shapes, as most of the existing approaches do, e.g. MPEG-7 visual descriptors [7], we broke contours into segments and describe curve contour geometry, individually and in relation with neighbor contours.

The main contribution of our work is however the combination of the proposed descriptors, which together form a highly descriptive feature set that is especially well-suited for video genre classification. The remainder of the paper is organized as follows: Section 2, Section 3, Section 4 and Section 5 deal with feature extraction: audio, temporal, color and contour, respectively. Experimental results are presented in Section 6 while Section 7 presents the conclusions and discuses future work.

2 Audio Descriptors

Audio information is an important cue when addressing automatic genre classification. Most of the common video genres have very specific audio signatures, e.g. music clips contain music, in news there are a lot of monologues/dialogues, documentaries have a mixture of natural sounds, speech and ambience 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 [8]. The proposed set of audio descriptors, called block-level audio features, have the key advantage of capturing also local temporal information from the audio track. Temporal integration is realized by analyzing sequences of consecutive frames *called blocks*, in a time-frequency representation, instead of using single frames only. Blocks are of variable length and can be overlapping (e.g. by 50% of their frames). After converting the video soundtrack into a $22kHz$ mono signal, we compute short-time Fourier transform and perform a mapping of the frequency axis according to the logarithmic cent-scale to account for the logarithmic human frequency perception. Then, the following complex audio features are derived:

Spectral Pattern (SP)**:** characterize the soundtrack's timbre via modeling those frequency components that are simultaneously active. Dynamic aspect of the signal are kept by sorting each frequency band of the block along the time

axis. The block width varies depending on the extracted patterns, which allows to capture temporal information over different time spans.

Delta Spectral Pattern (DSP)**:** captures the strength of onsets. To emphasize onsets, first the difference between the original spectrum and a copy of the original spectrum delayed by 3 frames is computed. Then, each frequency band is sorted along the time axis similar to the spectral pattern.

Variance Delta Spectral Pattern ($VDSP$)**:** is basically an extension of the delta spectral pattern and captures the variation of the onset strength over time.

Logarithmic Fluctuation Pattern (LFP)**:** captures the rhythmic aspects of the audio signal. In order to extract the amplitude modulations out of the temporal envelope in each band, periodicities are detected by computing the FFT along each frequency band of a block.

Spectral Contrast Pattern (SCP)**:** roughly estimates the *"tone-ness"* of an audio track. For each frame, within a block, the difference between spectral peaks and valleys in 20 sub-bands is computed and the resulting spectral contrast values are sorted along the time axis in each frequency band.

Correlation Pattern (CP)**:** To capture the temporal relation of loudness changes over different frequency bands, the correlation coefficient among all possible pairs of frequency bands within a block is used. The resulting correlation matrix forms the so-called correlation pattern.

These audio features in combination with a Support Vector Machine (SVM) classifier constitute a highly efficient automatic music classification system. During the last run of the Music Information Retrieval Evaluation eXchange, this approach ranked first with respect to the task of automatic music genre classification [8]. However, the proposed approach has not yet been applied to automatic video genre classification. Existing approaches are limited to use standard audio features, e.g. a common approach is to use Mel-Frequency Cepstral Coefficients (MFCC) or to compute time domain features, e.g. Root Mean Square of signal energy (RMS), or Zero-Crossing Rate (ZCR) [2] (preliminary tests proved the superiority of the block-based representation over classic MFCC features).

3 Temporal Structure Descriptors

As stated in the Introduction, temporal descriptors are derived using a classic confirmed approach, i.e. analyzing the frequency of shot changes [4]. Compared to existing approaches, we determine the action content based on human perception. Temporal based information is strongly related to movie genre, e.g. commercials and music clips tend to have a high visual tempo, commercials use a lot of gradual transitions, documentaries have a reduced action content, etc.

One of the main success factors of temporal descriptions is an accurate preceding temporal segmentation. To this end we detect both cuts and also gradual transitions. Cuts are detected using an adaptation of the histogram-based approach proposed in [9]. Fades and dissolves are detected using a pixel-level

statistical approach [10] and the analysis of fading-in and fading-out pixels (adaptation of [11]), respectively. Then, the temporal descriptors are computed, thus:

Rhythm. To capture the movie's visual changing tempo, first we compute the relative number of shot changes occurring within a time interval $T = 5s$, denoted ζ_T. Then, the rhythm is defined as the movie average shot change ratio, $E\{\zeta_T\}$.

Action. We aim at highlighting two opposite situations: video segments with a high action content (denoted hot action, e.g. fast changes, fast motion, visual effects, etc.) with $\zeta_T > 3.1$, and video segments with low action content (i.e. containing mainly static scenes) with $\zeta_T < 0.6$. Thresholds were determined experimentally. Several persons were asked to manually label video segments into the previous two categories. Based on this ground truth, we determined the average ζ_T intervals for each type of action content. Further, we quantify the action content with two parameters, hot-action ratio (HA) and low-action ratio (LA): $HA = T_{HA}/T_{total}$, $LA = T_{LA}/T_{total}$, where T_{HA} and T_{LA} represent the total length of hot and low action segments, respectively, and T_{total} is the movie total length.

Gradual Transition Ratio. High amounts of gradual transitions are in general related to a specific video contents, therefore we compute: $GT = (T_{dissolves} + T_{fade-in} + T_{fade-out})/T_{total}$, where T_X represents the total duration of all the gradual transitions of type X. This provides information about editing techniques which are specific to certain genres, like movies or artistic animated movies.

4 Color Descriptors

Color information is an important source to describe visual content. Most of the existing color-based genre classification approaches are limited to use intensity-based parameters or generic low-level color features, e.g. average color differences, average brightness, average color entropy, variance of pixel intensity, standard deviation of gray level histograms, percentage of pixels having saturation above a certain threshold, lighting key (measures how well light is distributed), object color and texture, etc. [2].

We propose a more elaborated strategy which addresses the perception of the color content [12]. One simple and efficient way to accomplish this is with the help of color names; associating names with colors allows everyone to create a mental image of a given color or color mixture. We project colors on to a color naming system and colors properties are described using: statistics of color distribution, elementary hue distribution, color visual properties (e.g. amount of light colors, warm colors, saturated colors, etc.) and relationship of color (adjacency and complementarity).

Our strategy is motivated by the fact that different genres have different global color signatures, e.g. animated movies have specific color palettes and color contrasts (light-dark, cold-warm), music videos and movies tend to have darker colors (mainly due to the use of special effects), sports usually show a predominant hue (e.g. green for soccer, white for ice hockey), and so on.

Prior to parameter extraction, we use an error diffusion scheme to project colors into a more manageable color palette, i.e. the non-dithering 216 color Webmaster palette (which provides an efficient color naming system). Further, the proposed color parameters are computed as follows:

Global Weighted Color Histogram is computed as the weighted sum of each shot color histogram, thus: $h_{GW}(c) = \sum_{i=0}^{M} \left[\frac{1}{N_i} \sum_{j=0}^{N_i} h^j_{shot_i}(c) \right] \cdot \frac{T_{shot_i}}{T_{total}}$, where M is the total number of video shots, N_i is the total number of the retained frames for the shot i (we use temporal sub-sampling), $h^j_{shot_i}$ is the color histogram of the frame j from the shot i, c is a color index from the Webmaster palette (we use color reduction) and T_{shot_i} is the length of the shot i. The longer the shot, the more important the contribution of its histogram to the movie's global histogram.

Elementary Color Histogram. The next feature is the distribution of elementary hues in the sequence, thus: $h_E(c_e) = \sum_{c=0}^{215} h_{GW}(c)|_{Name(c_e) \subset Name(c)}$, where c_e is an elementary color from the Webmaster color dictionary (colors are named according to color hue, saturation and intensity) and $Name()$ returns a color's name from the palette dictionary.

Color Properties. With this feature set we aim at describing color properties. We define several color ratios. For instance, light color ratio, P_{light}, reflects the amount of bright colors in the movie, thus: $P_{light} = \sum h_{GW}(c)|_{W_{light} \subset Name(c)}$, where c is a color with the property that its name contains one of the words defining brightness, i.e. $W_{light} \in \{$"light", "pale", "white"$\}$. Using the same reasoning and keywords specific to each property, we define dark color ratio (P_{dark}), hard saturated color ratio (P_{hard}), weak saturated color ratio (P_{weak}), warm color ratio (P_{warm}) and cold color ratio (P_{cold}). Additionally, we capture movie color wealth with two parameters: color variation, P_{var}, which accounts for the amount of significant different colors and color diversity, P_{div}, defined as the amount of significant different color hues.

Color Relationship. Finally, we compute P_{adj}, the amount of similar perceptual colors in the movie and P_{compl}, the amount of opposite perceptual color pairs.

5 Contour Descriptors

The last category of descriptors provide information based on visual structures, that is on contours and their relations. So far, contour information was only limitedly exploited within genre classification. For instance, some approaches use MPEG-7 inspired contour descriptors [7], e.g. use of texture orientation histograms, edge direction histograms, edge direction coherence, which are highly low-level edge pixel statistics.

Our approach in contrast, proposes a novel method which uses curve partitioning and curve description [13]. The contour description is based on a characterization of geometric attributes for each individual contour, e.g. degree of

curvature, angularity, "wiggliness", and so on. These attributes are taken as parameters in a high-dimensional image vector and have been exploited in a (statistical) classification task with good success. For instance, the system has achieved the benchmark in the photo-annotation task of the ImageCLEF competition 2010 where this approach ranks in the upper third of all performances.

Contour Characterization. Contour processing starts with edge detection, which is performed with the Canny edge detection algorithm. 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; abbreviated c_m, c_s respectively) and the "fuzziness", obtained from the convolution of the image with a blob filter (f_m, f_s, respectively).

Pair Relations. In addition to the attributes for individual contours, we also obtain attributes for pairs of contours which are selected based on spatial proximity (i.e. either their contour endpoints are proximal or in the proximity of the other segment). For each selected pair, a number of geometric attributes are determined such as the angular direction of the pair, denoted γ_p; distance between the proximal contour end points, denoted d_c; distance between the distal contour end points, denoted d_o; distance between the center (middle) point of each segment, denoted d_m; average segment length, denoted l; symmetry of the two segments, denoted y; degree of bendness of each segment, denoted b_1 and b_2; structural biases, abbreviated with $\hat{s}$, that express to what degree the pair alignment is a L feature ($\hat{s}_L$), T feature ($\hat{s}_T$) or a "closed" feature (two curved segments facing each other as '( )', $\hat{s}_{()}$).

The structural information is extracted only from a summary of the movie. In this case, we retain around 1% of each shot frames (uniformly distributed). For each image, contour properties are captured with histograms. To address the temporal dimension, at sequence level, resulting feature vectors are averaged forming so the structure signature of the movie.

6 Experimental Results

To evaluate the descriptive power of the proposed audio-visual content descriptors we have selected seven of the most common video genres, namely: *animated movies*, *commercials*, *documentaries*, *movies*, *music videos*, *news broadcast* and *sports*. The data set consists of 30 sequences for each genre, summing up more then 91 hours of video footage. Video materials were retrieved from several TV programmes, thus: 20h30min of animated movies (long, short clips and series, sources: Folimage - France, Disney, Pixar and DreamWorks animation companies); 15min of commercials (source 1980th TV commercials and David Lynch clips); 22h of documentaries (wildlife, ocean, cities and history, source BBC,

IMAX, Discovery Channel); 21h57min of movies (long, episodes and sitcom, e.g. Friends, X-Files, Sex and the City series); 2h30min of music (pop, rock and dance video clips, source MTV Channel); 22h of news broadcast (source TVR Romanian National Television Channel); 1h55min of sports (various clips from the Internet). Prior to analysis, a basic normalization is adopted by converting all sequences to a reference video format.

For our classification experiments we have selected three binary classifiers, namely: K-Nearest Neighbors (KNN, with k=1, cosine distance and majority rule), Support Vector Machines (SVM, with a linear kernel) and Linear Discriminant Analysis (use PCA to reduce dimensionality). Method parameters were tuned based on preliminary experimentations. All evaluations are conducted using a cross-validation approach, i.e. generating all possible combinations between training and test data. Additionally, we vary the amount of training data (from 10% to 70%) and test different combination of descriptors.

To assess performance, at genre level we evaluate average precision (P) and recall (R) ratios, thus: $P = \overline{TP}/(\overline{TP} + \overline{FP})$, $R = \overline{TP}/(\overline{TP} + \overline{FN})$, where $\overline{TP}$, $\overline{FP}$ and $\overline{FN}$ represent the *average* number of true positives, false positives and false negatives, respectively, computed over all experimentations for a given amount of training data. As a global measure of performance we compute F_{score} ratio and average correct classification ($\overline{CD}$), thus: $F_{score} = 2 \cdot P \cdot R/(P + R)$, $\overline{CD} = \overline{N_{GD}}/N_{total}$, were $\overline{N_{GD}}$ is the average number of good classifications (in both classes, target and others) and N_{total} is the number of test sequences. Experimental results are presented in the following.

6.1 One Genre at a Time Classification

In Figure 1 we present average precision against recall for different amounts of training data and different descriptor combinations, as well as the overall correct detection $\overline{CD}$ (descriptors are combined based on early fusion). We obtain very promising results considering the content similarity of some of the genres and also compared to the literature (see Section 1). We obtain $P \in [87.5\%; 100\%]$ (from which $P > 95\%$ for music, news, commercials and sports), and $R \in [77.6\%; 100\%]$ (excluding animated movies and commercials, we achieve $R > 95\%$). At global level, the overall correct classification ratio ranges from 92.2% to 97.2% while the highest F_{score} is up to 90.6%. One may observe that the overall performance is high, even for a reduced amount of training data, thus $\overline{CD} > 92\%$ with only 10% of data as training data (i.e. from 189 sequences, in average 174 were correctly assigned to one of the two classes, target genre and others).

The most interesting result is however that each descriptor set highlights relatively different properties of the video contents, as the most efficient approach (both in terms of overall classification performance and genre precision and recall) is the combination of all audio-visual descriptors (i.e. audio-contour-color-action, see SVM results depicted with the red line in Figure 1). Table 1 summarizes the precision and recall in this case (these results are encircled in Figure 1).

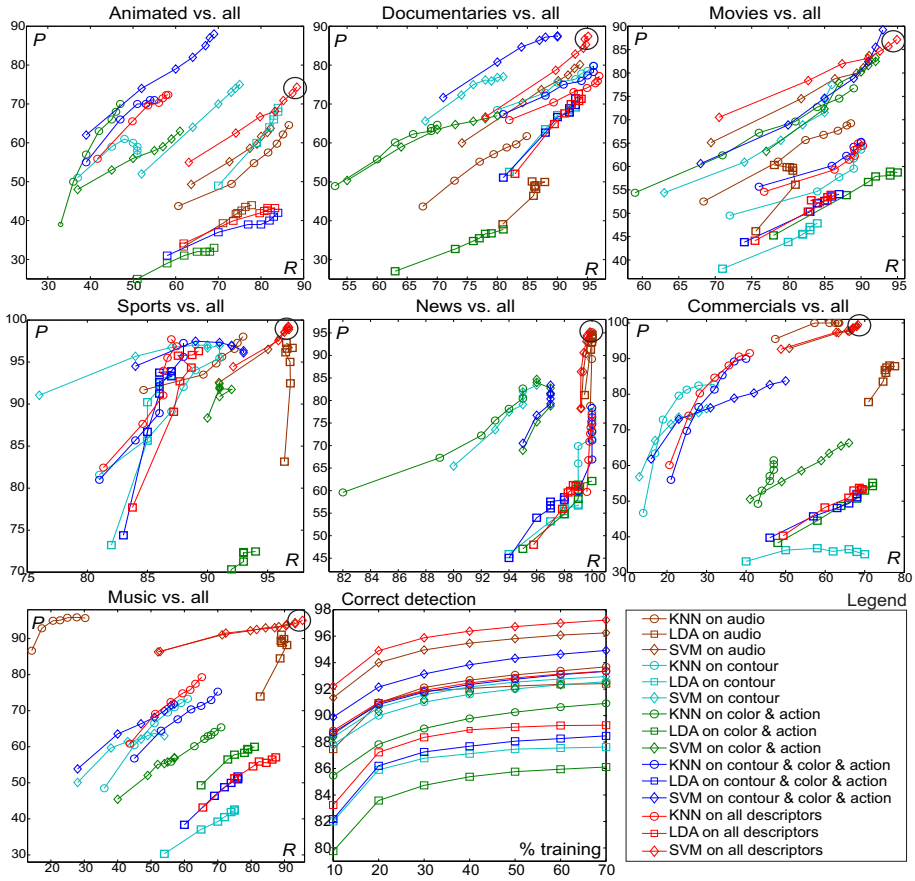

Fig. 1. Precision (P) against recall (R) for different runs and amounts of training data (increases along the curves from 10% to 70%) and overall correct classification ($\overline{CD}$)

Table 1. SVM vs. KNN and LDA (using all audio-visual descriptors)

genre	Precision (P)			Recall (R)		
	SVM	KNN	LDA	SVM	KNN	LDA
animated	**74.3%**	72.3%	43.2%	**88.4%**	58.2%	83.3%
documentaries	**87.4%**	77.2%	72.6%	**95.1%**	96.3%	93.5%
movies	**87.1%**	65%	53.9%	**94.9%**	89.6%	85.8%
music	**95.1%**	79.3%	57%	**95.4%**	65.2%	87.3%
sports	**99.3%**	97.7%	96.3%	**96.7%**	86.9%	89.2%
news	**95.2%**	76.9%	60.8%	**99.8%**	99.9%	99.1%
commercials	**99.5%**	91.5%	53.3%	**68.3%**	40.9%	69.4%

Globally, the lowest accuracy is obtained for animated movies and commercials, which is mainly due to their heterogenous contents and resemblance with other genres, e.g. many commercials include animation, music clips are similar to commercials, movies may contain commercial-like contents, etc. On the other hand, the best performance (as anticipated) is obtained for genres with a certain repetitiveness in content structure, i.e. news and sports (average precision or recall up to 100%).

In what concerns the informational sources, compared to visual information, audio information proves to be highly efficient to this task, alone leading to very good classification ratios (depicted with Maroon in Figure 1). At genre level, audio features are more accurate for classifying music, sports, news and commercials, which have specific audio patterns. On the other hand, contour and color-action information used alone, prove to be less efficient. Contour parameters, compared to color-action parameters, provide better performance for documentaries, sports and news, which have specific signatures, e.g. skyline contours, people silhouettes, etc. (depicted with Cyan in Figure 1). Compared to contours, color-action features perform better for music, commercials, movies and news (which can be assigned to the specific rhythm and color diversity, depicted with Green in Figure 1). Compared to audio, visual descriptors together are more discriminative for animated, movies and documentaries (depicted with Blue in Figure 1). As stated before, the best performance in classifying each individual genre is however achieved when using all audio-visual information.

6.2 Descriptor-Based Visualization

In our final experiment we try to find out whether the proposed features are discriminative enough to provide genre-based separation for real browsing applications. Movies were displayed on a 3D spherical coordinate system according to the first three principal components of the audio-visual descriptors, thus: inclination (θ) - 1st component (normalized in $[0;\pi]$), azimuth (φ) - 2nd component (normalized in $[0;2\pi]$) and radius (r) - 3rd component (normalized in $[0;1]$). Several screenshots taken from different angles are presented in Figure 2 (a demo is available at `http://imag.pub.ro/~bionescu/index_files/MovieGlobe.avi`).

Although, we use only the first three principal components (which account for up to 94% of the initial data variance), one may observe that certain genres are visibly grouping together, which is quite an interesting result. Due to the similarity of the content and structure, the mostly regrouped are the news (see view C) and sports (see view D). Other genres tend to be more "interleaved" (e.g. documentaries, see view B), which is at some point expectable, considering the fact that even for human observer is difficult to draw a sharp delimitation between genres. Nevertheless, sequences with similar contents tend to regroup around a basis partition (see in view A).

Enhanced by genre labeling provided by the SVM classification, this might be a powerful genre-based browsing tool. Even though this experiment proves the potential of our descriptors with real browsing applications, these are however preliminary results and more elaborated tests are to be conducted.

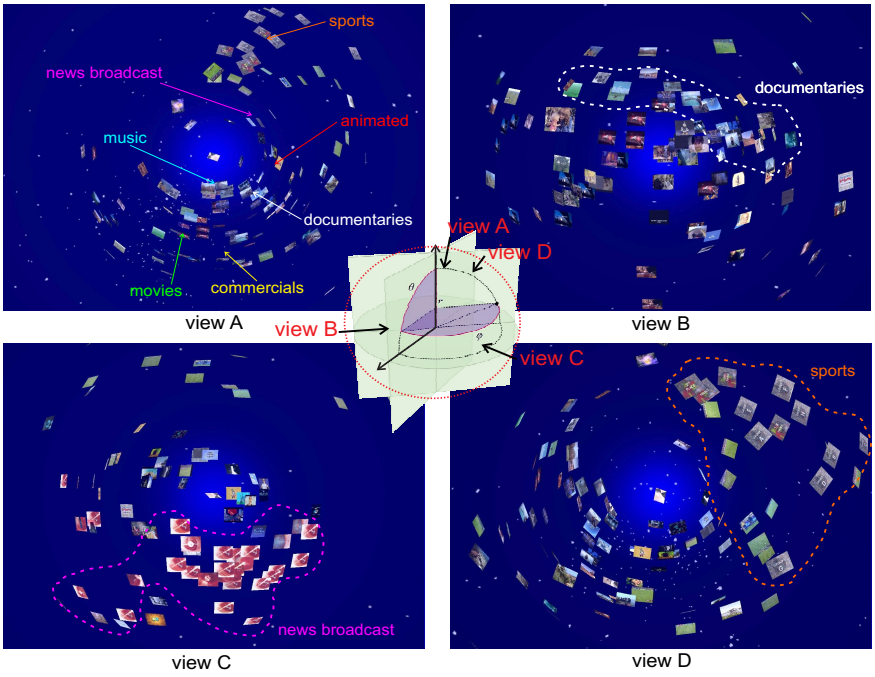

Fig. 2. Feature-based 3D movie representation (each movie is represented with one image). View A to D are screenshots made from different perspectives (the used points of view are synthesized with the system diagram presented in the center).

7 Conclusions

In this paper we addressed the issue of automatic video genre categorization and we have proposed four categories of content descriptors: block-level audio features, temporal structure-based action descriptors, perceptual color descriptors and contour statistics.

Although these sources of information have already been exploited in the literature, the main contribution of our work is the way we compute the content descriptors and the high descriptive power of the combination of these descriptors. An extensive evaluation was performed based on 91 hours of video footage. We achieve average precision and recall ratios within $[87\%-100\%]$ and $[77\%-100\%]$, respectively, while average correct classification is up to 97%.

Additionally, preliminary experiments based on a prototypical video browsing system demonstrate the prospective application potential of our approach. Future work will mainly focus on more detailed sub-genre classification and on extending the scope of our work towards web media platforms (e.g. blip.tv, see MediaEval campaign).

Acknowledgments. This work was supported by the Romanian Sectoral Operational Programme Human Resources Development 2007-2013 through the Financial Agreement POSDRU/89/1.5/S/62557 and by the Austrian Science Fund (FWF): L511-N15.

References

1. Smeaton, A.F., Over, P., Kraaij, W.: High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. In: Multimedia Content Analysis, Theory and Applications, pp. 151–174. Springer, Berlin (2009) ISBN 978-0-387-76567-9
2. Brezeale, D., Cook, D.J.: Automatic Video Classification: A Survey of the Literature. IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews 38(3), 416–430 (2008)
3. Roach, M.J., Mason, J.S.D.: Video Genre Classification using Dynamics. In: IEEE Int. Conf. on Acoustics, Speech and Signal Processing, USA, pp. 1557–1560 (2001)
4. Yuan, X., Lai, W., Mei, T., Hua, X.-S., Wu, X.-Q., Li, S.: Automatic Video Genre Categorization using Hierarchical SVM. In: IEEE Int. Conf. on Image Processing, pp. 2905–2908 (2006)
5. Montagnuolo, M., Messina, A.: Parallel Neural Networks for Multimodal Video Genre Classification. Multim. Tools and Applications 41(1), 125–159 (2009)
6. Wang, H., Divakaran, A., Vetro, A., Chang, S.-F., Sun, H.: Survey of Compressed-Domain Features used in Audio-Visual Indexing and Analysis. Journal of Visual Communication and Image Representation 14(2), 150–183 (2003)
7. Sikora, T.: The MPEG-7 Visual Standard for Content Description - An Overview. IEEE Trans. on Circ. and Systems for Video Technology 11(6), 696–702 (2001)
8. Seyerlehner, K., Schedl, M., Pohle, T., Knees, P.: Using Block-Level Features for Genre Classification, Tag Classification and Music Similarity Estimation. In: 6th Annual Music Information Retrieval Evaluation eXchange (MIREX 2010), Utrecht, Netherlands, August 9-13 (2010)
9. Ionescu, B., Buzuloiu, V., Lambert, P., Coquin, D.: Improved Cut Detection for the Segmentation of Animation Movies. In: IEEE Int. Conf. on Acoustic, Speech and Signal Processing, Toulouse, France (2006)
10. Fernando, W.A.C., Canagarajah, C.N., Bull, D.R.: Fade and Dissolve Detection in Uncompressed and Compressed Video Sequence. In: IEEE Int. Conf. on Image Processing, Kobe, Japan, pp. 299–303 (1999)
11. Ionescu, B., Buzuloiu, V., Lambert, P., Coquin, D.: Dissolve Detection in Abstract Video Contents. In: IEEE Int. Conf. on Acoustic, Speech and Signal Processing, Prague, Czech Republic (2011)
12. Ionescu, B., Coquin, D., Lambert, P., Buzuloiu, V.: A Fuzzy Color-Based Approach for Understanding Animated Movies Content in the Indexing Task. Eurasip Journal on Image and Video Processing (2008), doi:10.1155/2008/849625
13. Rasche, C.: An Approach to the Parameterization of Structure for Fast Categorization. Int. Journal of Computer Vision 87(3), 337–356 (2010)