A contour-color-action approach to automatic classification of several common video genres

Bogdan E. Ionescu[1,2], Christoph Rasche[1],
Constantin Vertan[1], and Patrick Lambert[2]

[1] LAPI - University Politehnica of Bucharest, 061071 Bucharest, Romania
[2] LISTIC - Polytech'Savoie, B.P. 806, 74016 Annecy, France
{bionescu,rasche,cvertan}@alpha.imag.pub.ro
patrick.lambert@univ-savoie.fr

Abstract. We address the issue of automatic video genre retrieval. We propose three categories of content descriptors, extracted at temporal, color and structural level. At temporal level, video content is described with visual rhythm, action content and amount of gradual transitions. Colors are globally described with statistics of color distribution, elementary hues, color properties and relationship. Finally, structural information is extracted at image level and histograms are built to describe contour segments and their relations. The proposed parameters are used to classify 7 common video genres, namely: animated movies/cartoons, commercials, documentaries, movies, music clips, news and sports. Experimental tests using several classification techniques and more than 91 hours of video footage prove the potential of these parameters to the indexing task: despite the similarity in semantic content of several genres, we achieve detection ratios ranging between $80 - 100\%$.

Keywords: video genre classification, action content, color properties, contour structural information, video indexing.

1 Introduction

An interesting challenge in the fields of content-based video indexing is the automatic cataloging of video footage into predefined semantic categories. This can be performed either globally, by classifying video into one of several main genres, e.g. cartoons, music, news, sports or even further into some sub-genres, e.g. identifying specific types of sports (football, hockey, etc.), movies (drama, thriller, etc.); or either locally by focusing on classifying segments of video such as retrieving concepts, e.g. outdoor vs. indoor, violence, action, etc. [1].

Being related to the issue of data mining, video genre classification involves two steps: *feature extraction* and *data classification*. Feature extraction and selection is one critical step towards the success of the classification task. The main challenge is to derive attributes discriminant enough to emphasize particularities of each genre while preserving a relatively reduced number of features. Most of the existing feature extraction approaches rely on visual elements, like color, temporal structure, objects, motion, etc., which are to be used either alone or in combination with text or audio features. A complete state-of-the art of the literature on this matter is presented in [2]. In the following we shall highlight several approaches we consider representative and related to our work.

For instance, [3] addresses the genre classification task using only video dynamics. Motion information is extracted at two levels: background camera motion and foreground or object motion and 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 simplicity of this single modal approach, it is able to achieve detection errors below 6%.

A more complex approach which uses spatio-temporal information is proposed in [4]. At temporal level, video content is described in terms of 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 movie, commercial, news, music and sports; 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%, for sports categorization it is 97% while for movie categorization it is around 81.3%, however no information is provided on the recall ratios.

An interesting true multi-modal approach, which combines several types of content descriptors, is proposed in [5]. Features are extracted from four informative sources, which include visual-perceptual information (colour, 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 this study we propose three categories of content descriptors derived at temporal, color and contour-based levels. Compared to existing literature, e.g. MPEG-7 descriptors, they have some advantages. Temporal descriptors (e.g. action) are determined based on the perception of action at different levels (user experiments have been conducted). Color descriptors involve also the perception of colors (transposed from the semantic analysis of artistic animated movies [6]) and color is generically described in terms of statistics of color distribution, elementary hues, color properties (e.g. amount of light colors, cold colors, etc.) and relationship of adjacency and complementarity. Contour descriptors focus not on closed shapes (although difficult to obtain) but propose to describe curved segments and contour geometry is described individually (e.g. orientation, degree of curvature, symmetry, etc.) or in relation with other neighboring contours (e.g. angular direction, geometric alignment, etc.). The method is transposed from static image indexing, where it has been successfully validated on retrieving tens of semantic concepts, e.g. outdoor, doors/entrances, fruits, people, etc. [7].

The main novel aspect is however the combination of all these parameters for the classification of 7 common genres. Each genre shows some specificity for these parameters (empirically determined), for instance: *animated movies/cartoons* - have particular color properties; *documentaries* - skyline contours are predominant, rhythm is rather slow; *music clips* - high visual rhythm, high action, darker color palette; *news broadcast* - people/face silhouettes are predominant; *commercials* - high rhythm, rather abstract like animated movies; *movies* - homogenous color palette, similar global rhythm, characters/faces occurrence is high, darker colors and *sports* - have few predominant hues, people silhouettes are predominant (see Section 5.1). Exhaustive experimental tests have been conducted on more than 91 hours of video footage and classification is performed using SVM (Support Vector Machines), KNN (K-Nearest Neighbor) and LDA (Linear Discriminant Analysis). Despite the difficulty of this task due to resemblance between several genres (e.g. music clips and commercials, movies and documentaries) the proposed parameters achieve average precision and recall ratios up to 97% and 100%, respectively.

The remainder of this paper is organized as follows: Section 2, Section 3 and Section 4 deal with feature extraction: temporal structure (action), color properties and image structural information (contour), respectively. Experimental results are presented in Section 5 while Section 6 presents the conclusions and discuses future work.

2 Action descriptors

The first feature set aims to capture the movie temporal structure in terms of *visual rhythm*, *action* and *amount of gradual video transitions*. These parameters are strongly related to movie contents, e.g. music clips have a high visual tempo, documentaries a low action content, commercials a high amount of gradual transitions, etc. The approach is based on some previous work [9]. It is carried out by first performing movie temporal segmentation, which roughly means the detection of video transitions. We detect cuts and two of the most frequent gradual transitions, i.e. fades and dissolves. Cut detection is performed using an adaptation of the histogram-based approach proposed in [10], while fade and dissolve detection are carried out using a pixel-level statistical approach [11] and the analysis of fading-in and fading-out pixels [12], respectively. Further, we determine the following parameters (see also Figure 1):

Rhythm. To capture the movie's changing tempo, we define first a basic indicator, denoted $\zeta_T(i)$, which represents the relative number of shot changes occurring within the time interval of T seconds, starting from a frame at time index i ($T = 5s$, experimentally determined). Based on ζ_T, we define the movie rhythm as the movie's average shot change speed, $\bar{v}_T$, i.e. the average number of shot changes over the time interval T for the entire movie [8], thus:

$$\bar{v}_T = E\{\zeta_T(i)\} = \sum_{t=1}^{T\cdot 25} t \cdot f_{\zeta_T(i)}(t) \tag{1}$$

in which $T \cdot 25$ represents the number of frames of the time window (at 25 fps) and $f_{\zeta_T(i)}$ is the probability density of $\zeta_T(i)$ given by:

$$f_{\zeta_T(i)}(t) = \frac{1}{N_T} \sum_{i \in W_T} \delta(\zeta_T(i) - t) \tag{2}$$

in which N_T is the total number of time windows of size T seconds (defining the set W_T), i is the starting frame of the current analyzed time window and $\delta(t) = 1$ if $t = 0$ and 0 otherwise.

Action. To determine the following parameters, we use the basic assumption that, in general, action content is related to a high frequency of shot changes [13]. We aim at highlighting two opposite situations: video segments with a high action content (hot action) and video segments with low action content [8].

First, at a coarse level, we highlight segments which show high number of shot changes ($\zeta_T > 2.8$), i.e. candidates for hot action, and a reduced number of shot changes ($\zeta_T < 0.71$), i.e. low action. Thresholds were set experimentally after manually analyzing ζ_T values for several representative action segments of each class (adaptation of [8]). To reduce over-segmentation of action segments, we merge neighboring action segments at a time distance below T seconds (the size of the time window). Further, we remove unnoticeable and irrelevant action segments by erasing small action clips less than the analysis time window T. Finally, all action clips containing less than $N_s = 4$ video shots are being removed. Those segments are very likely to be the result of false detections, containing one or several gradual transitions (e.g. a "fade-out" - "fade-in" sequence).

Based on this information, action content is described with two parameters, hot-action ratio (HA) and low-action ratio (LA):

$$HA = \frac{T_{HA}}{T_{total}}, \quad LA = \frac{T_{LA}}{T_{total}} \tag{3}$$

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. The last parameter is related to the amount of the gradual transitions used within the movie. We compute the gradual transition ratio (GT):

$$GT = \frac{T_{dissolves} + T_{fade-in} + T_{fade-out}}{T_{total}} \tag{4}$$

where T_x represents the total duration of all the gradual transitions of type x.

3 Color descriptors

The next feature set aims to capture the movie's global color contents in terms of statistics of color distribution, elementary hues, color properties and color relationship. This is carried out using an adaptation of the approach proposed in [6]. Prior to the analysis, several pre-processing steps are adopted. To reduce complexity, color features are computed on a summary of the initial video. Each video shot is summarized by retaining only $p = 10\%$ of its frames as a subsequence centered with respect to the middle of the shot (experimental tests proved that 10% is enough to preserve a good estimation of color distribution). The retained frames are down-sampled to a lower resolution (e.g. average width around 120 pixels). Finally, true color images are reduced to a more convenient color palette. We have selected the non-dithering 216 color Webmaster palette due to its consistent color wealth and the availability of a color naming system. Color mapping is performed using a minimum L*a*b* Euclidean distance approach applied using a Floyd-Steinberg dithering scheme [14]. The proposed color parameters are determined as follows (see also Figure 1).

Global weighted color histogram captures the movie's global color distribution. It is computed as the weighted sum of each individual shot average color histogram, thus:

$$h_{GW}(c) = \sum_{i=0}^{M} \left[\frac{1}{N_i} \sum_{j=0}^{N_i} h^{j}_{shot_i}(c) \right] \cdot \frac{T_{shot_i}}{T_{total}} \tag{5}$$

where M is the total number of video shots, N_i is the total number of the retained frames from the shot i (i.e. $p = 10\%$), $h^{j}_{shot_i}()$ is the color histogram of the frame j from shot i, c is a color index from the Webmaster palette and T_{shot_i} is the total length of the shot i. The longer the shot, the more important the contribution of its histogram to the movie's global histogram. Defined in this way, values of $h_{GW}()$ account for the global color apparition percentage in the movie (values are normalized to 1, i.e. a frequency of occurrence of 100%).

Elementary color histogram. The next feature is the elementary color distribution which is computed, thus:

$$h_E(c_e) = \sum_{c=0}^{215} h_{GW}(c)|_{Name(c_e) \subset Name(c)} \tag{6}$$

where c_e is an elementary color from the Webmaster color dictionary (colors are named according to the color's hue, saturation and intensity), $c_e \in \Gamma_e$ with $\Gamma_e = \{$"Orange", "Red", "Pink", "Magenta", "Violet", "Blue", "Azure", "Cyan", "Teal", "Green", "Spring", "Yellow", "Gray", "White", "Black"$\}$ and $Name()$ returns a color's name from the palette dictionary. In this way, each available color is projected in $h_E()$ on to its elementary hue, therefore disregarding the saturation and intensity information. This mechanism assures invariance to color fluctuations (e.g. illumination changes).

Color properties. The next parameters aim at describing, first, color perception by means of light/dark, saturated/non-saturated, warm/cold colors and second, color wealth by quantifying color variation/diversity. Using previously determined histogram information in conjunction with color naming dictionary we define several color ratios.

For instance, light color ratio, P_{light}, which reflects the amount of bright colors in the movie, is computed thus:

$$P_{light} = \sum_{c=0}^{215} h_{GW}(c)|_{W_{light} \subset Name(c)} \tag{7}$$

where c is the index of a color with the property that its name (provided by $Name(c)$) contains one of the words defining brightness, i.e. $W_{light} \in \{$"light", "pale", "white"$\}$. Using the same reasoning and keywords specific to each color property, we define:

- dark color ratio, denoted P_{dark}, where $W_{dark} \in \{$"dark", "obscure", "black"$\}$;
- hard color ratio, denoted P_{hard}, which reflects the amount of saturated colors. $W_{hard} \in \{$"hard", "faded"$\} \cup \Gamma_e$, where Γ_e is the elementary color set (see equation 6, elementary colors are 100% saturated colors);
- weak color ratio, denoted P_{weak} which is opposite to P_{hard}, $W_{weak} \in \{$"weak", "dull"$\}$;
- warm color ratio, denoted P_{warm}, which reflects the amount of warm colors; in art, some hues are commonly perceived to exhibit some levels of warmth, namely: "Yellow", "Orange", "Red", "Yellow-Orange", "Red-Orange", "Red-Violet", "Magenta", "Pink" and "Spring";
- cold color ratio, denoted P_{cold}, where "Green", "Blue", "Violet", "Yellow-Green", "Blue-Green", "Blue-Violet", "Teal", "Cyan" and "Azure" are reflecting coldness.

Further, we capture movie color wealth with two parameters. Color variation, P_{var}, which accounts for the amount of significant different colors, is defined thus:

$$P_{var} = \frac{Card\{c|h_{GW}(c) > \tau_{var}\}}{216} \tag{8}$$

where c is a color index, h_{GW} is the global weighted histogram defined in equation 5 and $Card()$ is the cardinal function which returns the size of a data set. We consider a color significant enough for the movie's color distribution if it has a frequency of occurrence of more than 1% (i.e. $\tau_{var} = 0.01$). Color diversity, P_{div}, which reflects the amount of significant different color hues is defined on the elementary color histogram h_E using the same principle.

Color relationship. The final two parameters are related to the concept of perceptual relation of color in terms of adjacency and complementarity. Hence,

P_{adj} reflects the amount of similar perceptual colors in the movie (neighborhood pairs of colors on a perceptual color wheel, e.g. Itten's color wheel) and P_{compl} reflects the amount of opposite perceptual color pairs (antipodal).

4 Contour descriptors

The final descriptor set provides structural information in terms of contours and their relations. Contours are partitioned and represented as described in [7]. Similar to color information, contour information is extracted not from the entire movie but from a summary of the movie. In this case, we aim at retaining around 100 images evenly distributed with respect to video transitions. Retained frames are down-sampled to a lower resolution, whereby maintaining the image's aspect ratio (e.g. average width around 120 pixels). Contour processing starts with edge detection, which is performed with a Canny edge detector [16], and continues with creation of the local/global space (LG) for each extracted contour, followed by contour partitioning, segment extraction and finally contour description. To capture all relevant structural information, contours are extracted at several spatial scales (e.g. σ=1,2,3,5, see [16]). At the beginning, descriptors are determined for all four scales but later reduced by keeping only the most symmetric and smooth contours.

Contour signatures. A contour is iterated with a window (fixed chord: ω) which classifies a segment into three labels: bow, inflexion and straight, and additionally determines the amplitude of the segment. For a given window size, this leads to the "bowness" $\beta(v)$, inflexion $\tau(v)$ and straightness signature $\gamma(v)$, where v represents the arc length variable. For a range of window sizes, this leads to a set of signatures which describe the LG space, one for bows, $\beta_\omega(v)$, one for inflexions, $\tau_\omega(v)$, and one for straightness, $\gamma_\omega(v)$. The straightness signature is suppressed (γ set to 0) if at the same location a positive bowness value is present in the same or any higher window level ω.

Contour properties. Further, contours are partitioned at U turns, i.e. sharp curvatures of 180 degrees, which can be located in the bowness space $\beta_\omega(v)$. After application of this rule, any contour appears either as elongated in a coarse sense or as an arc. A contour is thus soft-classified as "wiggly" and "arced" by setting a scalar value that expresses the strength of these aspects (w and a respectively; if both values are 0 the contour is straight). From the wiggly contours, long straight and arced segments are extracted, as they could be locally grouped with other neighboring contours. Other geometric aspects that are derived from the LG space are:

- degree of curvature, denoted b;
- degree of circularity, denoted ζ (for arcs larger than 180 degrees);
- edginess parameter, denoted e, that expresses the sharpness of a curve (L feature or bow);
- symmetry parameter, denoted y, that expresses the "eveness" of the contour.

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) [7].

Contour relationship. Contour segments are then grouped, firstly as pairs. For each contour, three neighboring segments are searched (that potentially constitute useful pairs for description): one for each endpoint and one for its center point that forms a potential pair of parallel segments. The selection of appropriate pairs is based on a number of criteria and measures such as the spatial proximity between (proximal) endpoints, the structural similarity of the two segments and the symmetry of their alignment. Selected pairs are then geometrically described by the following dimensions:

- 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 '( )', $\hat{s}_{()}$).

In summary, at the image level, structural information is represented in a statistical manner using histograms. For each descriptor parameter, a 10-bin histogram is generated. The histograms are then concatenated to form a single descriptor vector. At movie level, feature vectors are averaged, forming so the structure signature of the movie.

5 Experimental results

To test the discriminative power of the proposed parameters in video genre classification, we have selected 7 of the most common genres, namely: *animated movies*, *commercials*, *documentaries*, *movies*, *music videos*, *news broadcast* and *sports*. Each genre is represented with 30 sequences recorded from several TV programmes, summing up to 210 sequences and more than 91 hours of video footage, thus: 20h30min of animated movies (long, short clips and series), 15min of commercials, 22h of documentaries (wildlife, ocean, cities and history), 21h57min of movies (long, episodes and sitcom), 2h30min of music (pop, rock and dance video clips), 22h of news broadcast and 1h55min of sports (mainly soccer).

5.1 Parameter examples

In Figure 1 and 2 we present average color (see Section 3), action (see Section 2) and contour (see Section 4) feature vectors for each of the 7 genres.

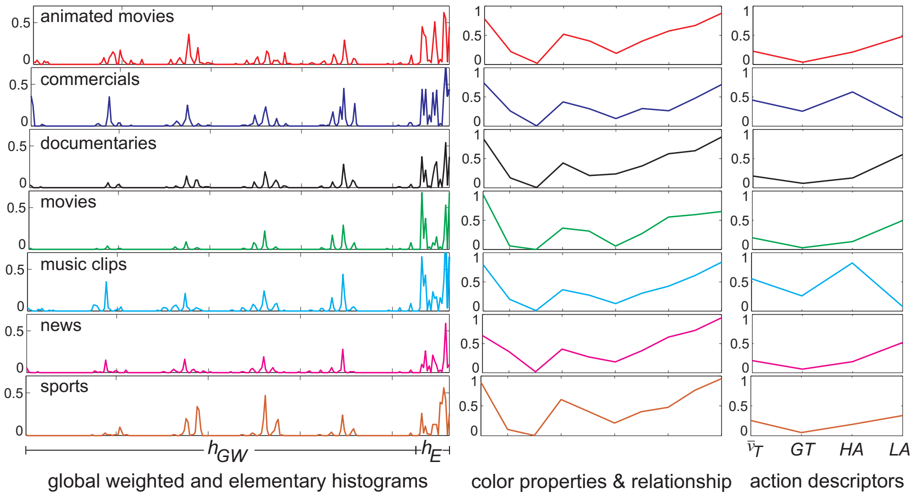

Fig. 1. Average color-action feature vectors for each genre.

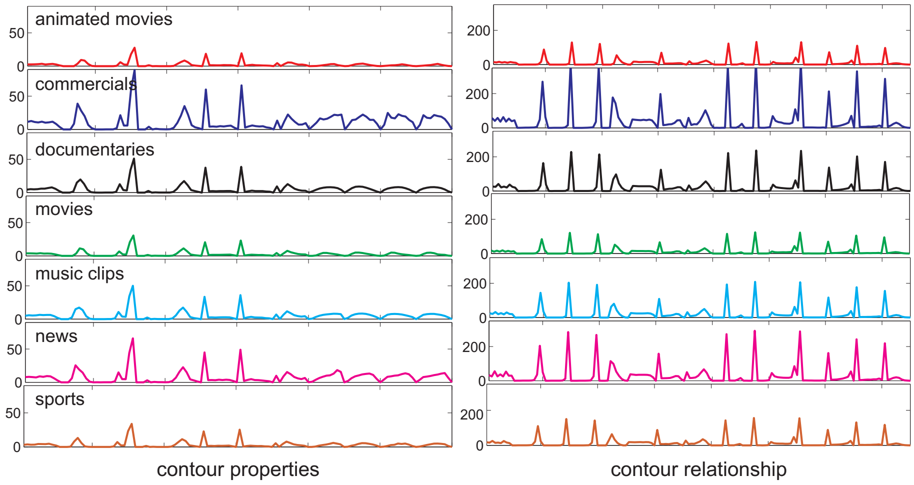

Fig. 2. Average contour feature vectors for each genre.

In average, the proposed parameters show a different comportment for each genre. For instance, commercials and music clips have a high visual rhythm and action content (see $\bar{v}_T$ and HA in Figure 1), animated movies have a different color pattern (more colors are being used, see h_{GW}) and most of the hues are used in important amounts (see h_E), movies and documentaries tend to have a reduced action content, sports have a predominant hue (see the predominant peak in h_E), commercials show an important symmetry of contours (see contour relationship in Figure 2), and so on. Discriminant power of the features is evidenced however in the classification task below.

5.2 The classification approach

Genre retrieval is carried out with a binary classification approach, i.e. one genre at a time vs. all others. We test several supervised classification methods, namely: the K-Nearest Neighbors algorithm (KNN, with k=1, cosine distance and majority rule), Support Vector Machines (SVM, with a linear kernel) and Linear Discriminant Analysis (LDA, with linear discriminant function applied on a PCA-reduced feature space) [15]. The method parameters were set to optimal values for this scenario after several tests.

As the choice of the training set may distort the accuracy of the results, we have adopted an exhaustive testing, i.e. training sequences are selected randomly and each classification is repeated over 1000 times in order to extract all possible combinations. Also, tests were performed for different amounts of training data, as depicted in Table 1.

Table 1. Training sets from 210 test sequences.

% training data	10%	20%	30%	40%	50%	60%	70%
total nb. of training sequences	21	42	63	84	105	126	147
(from which) # of current genre:	3	6	9	12	15	18	21
total nb. of test sequences	189	168	147	126	105	84	63
(from which) # of current genre	27	24	21	18	15	12	9

To assess performance we adopt several strategies. First, we evaluate average precision (P) and recall (R) ratios for each target class, thus:

$$P = \frac{\overline{TP}}{\overline{TP} + \overline{FP}}, \quad R = \frac{\overline{TP}}{\overline{TP} + \overline{FN}} \tag{9}$$

where $\overline{TP}$, $\overline{FP}$ and $\overline{FN}$ represent the *average* number of good detections (true positives), false detections (false positives) and non detections (false negatives), respectively. Secondly, to provide a global measure of performance we compute the average correct detection ratio, denoted $\overline{CD}$, and F_{score} ratio, thus:

$$\overline{CD} = \frac{\overline{N_{GD}}}{N_{total}}, \quad F_{score} = 2 \cdot \frac{P \cdot R}{P + R} \tag{10}$$

were $\overline{N_{GD}}$ is the average number of good classifications (for both classes, target genre and others) and $N_{total} = 210$ is the number of test sequences.

5.3 Discussion on precision and recall

Despite the strong semantic resemblance between different genres, the proposed parameters achieve good classification results. Figure 3 depicts the precision vs. recall curves for different amounts of training data (see Table 1) and an average Fscore ($\overline{F^{g}_{score}}$) over all genres. For all genres we achieve detection ratios above 80%, while for some particular genres detection ratios are close to 100%.

Due to the similarity of the content, the weakest classification performance is obtained for music and commercials, thus:

- for *music* the highest accuracy is obtained with LDA using color-action ($R =$ 81%) and the lowest false positive rate with KNN using contour-color-action ($P = 75\%$);
- for *commercials* the highest accuracy is obtained with LDA using color-action ($R = 72\%$) and the lowest false positive rate with KNN using contour-color-action ($P = 89\%$).

The high diversity of video material (short clips, long movies, episodes, cartoons, artistic animated movies) situates *animated movies* on an average classification performance, thus the highest accuracy is obtained with LDA using contour-color-action ($R = 84\%$) while the lowest false positives rate with SVM using contour-color-action ($P = 88\%$).

A relatively high classification accuracy is obtained for genres which show at some level a certain homogeneity in structure and content, namely: documentaries, movies, news and sports:

- for *documentaries* the highest accuracy and the lowest false positives rates are both obtained with KNN using contour-color-action ($R = 96\%$, $P =$ 80%);
- for *movies* the highest accuracy is obtained with LDA using color-action ($R = 95\%$) while the lowest false positives rate is obtained with SVM using contour-color-action ($P = 87\%$);
- for *news* the highest accuracy is obtained with KNN using contour-color-action ($R = 100\%$), as well as with KNN using contour and LDA using color-action (however, precision is lower for the last two), while the lowest false positives rate is obtained with SVM using color-action ($P = 85\%$);
- for *sports* the highest accuracy is obtained with LDA using color-action ($R = 94\%$) while the lowest false positives rate is obtained with KNN using contour-color-action ($P = 97\%$).

In general, detection ratios increase with the amount of training data (see Figure 3). Also, one may observe that the best performance tends to be achieved with the fusion of contour, color and action parameters.

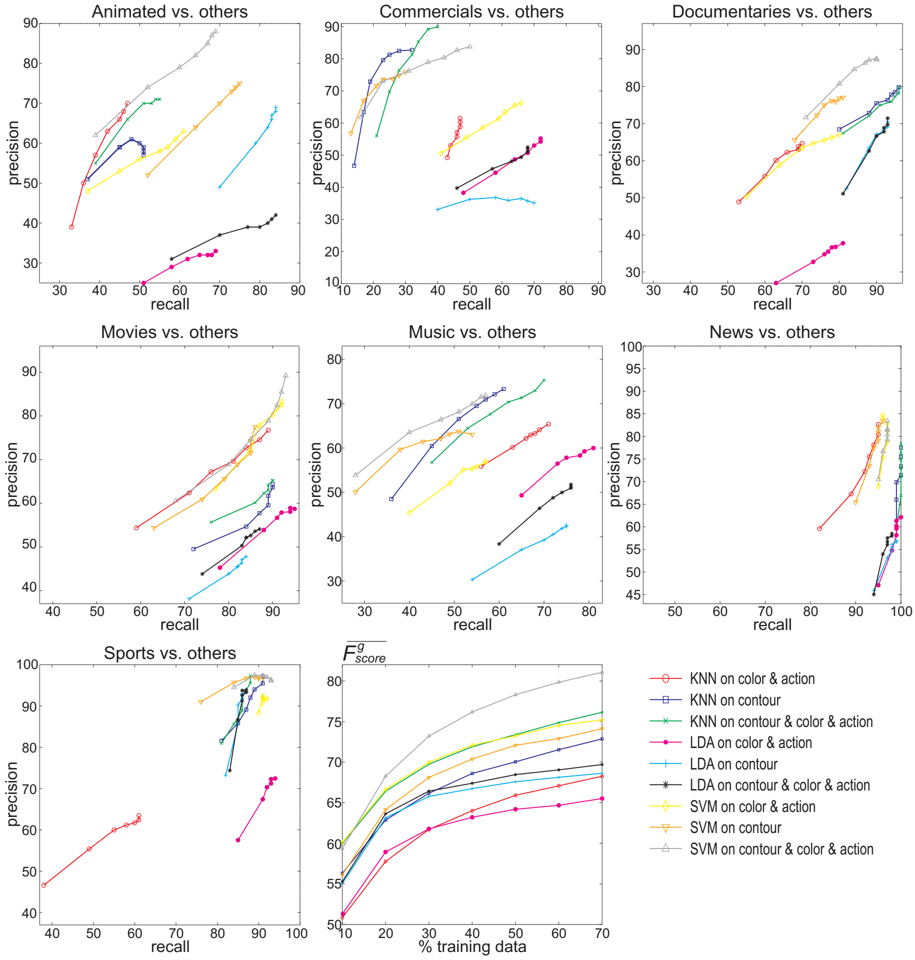

Fig. 3. Precision vs. recall curves for different runs and amounts of training data (% of training is increasing along the curves). $\overline{F^g_{score}}$ represents the average Fscore measure achieved for all genres.

5.4 Discussion on global Fscore and correct detection ratio

To asses overall performance, we use an average Fscore measure ($\overline{F^g_{score}}$, see Figure 3) and the average correct detection ratio ($\overline{CD}$, see Figure 4) which are computed over all genres (we use averaging as each genre is represented with equal number of sequences). Based on this information, the most powerful approach proves to be, again, the combination of all three feature classes, i.e. contour, color and action.

It provides the best result with SVM and KNN classifiers, namely: SVM - $\overline{F^g_{score}} = 81.06\%$, $\overline{CD} = 94.92\%$, followed by KNN - $\overline{F^g_{score}} = 76.14\%$, $\overline{CD} =$

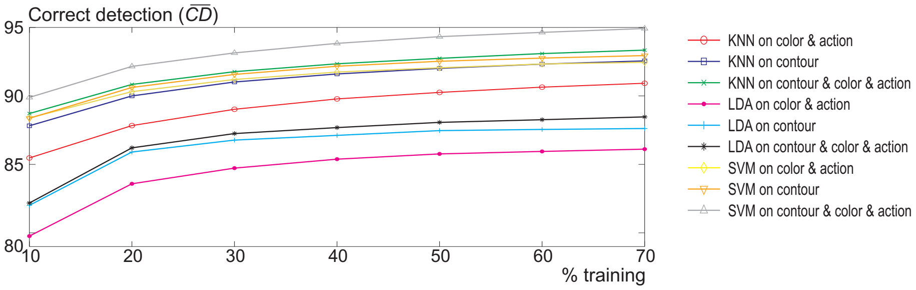

Fig. 4. Average correct detection ratio for all genres ($\overline{CD}$, the oX axis represents the amount of training data, see Table 1).

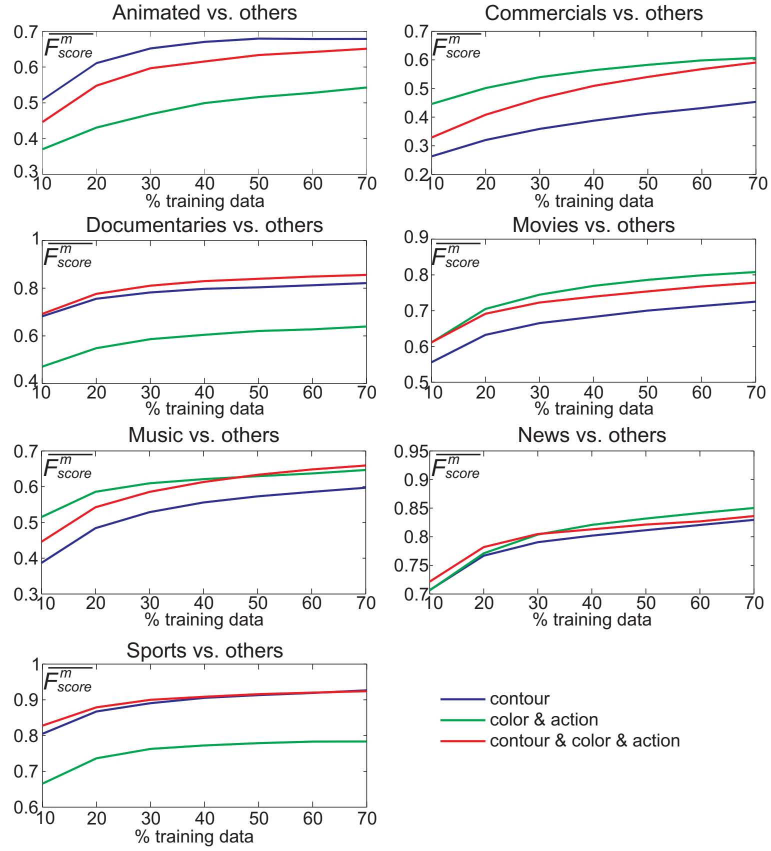

Fig. 5. Average Fscore measure per feature set and for all methods (KNN + SVM + LDA, $\overline{F^m_{score}}$).

93.34%. Also, SVM provides the highest accuracy for the smallest training set (see Figure 4), e.g. $\overline{CD} = 89.89\%$ (training 10%, see Table 1) or $\overline{CD} = 92.15\%$ (training 20%).

The final test was to determine the overall discriminant power of each feature set. For that, we compute an average Fscore for all three methods (KNN + SVM + LDA), denoted $\overline{F^m_{score}}$. The results are presented in Figure 5. Although this tends to be a relatively subjective evaluation, being related to the performance of the selected classification methods, we obtain some interesting results. Contour parameters, compared to color-action parameters, preserve their efficiency in retrieving specific object shapes, as proved for static images [7]. They provide the highest score for documentaries (skyline contours are frequent, $\overline{F^m_{score}} = 82.12\%$), sports (people silhouettes are frequent, $\overline{F^m_{score}} = 92.42\%$) and good results for news (anchorperson bust silhouette and people silhouette are frequent, $\overline{F^m_{score}} = 82.96\%$). On the other hand, compared to contours, color-action features perform better for music ($\overline{F^m_{score}} = 64.66\%$), commercials ($\overline{F^m_{score}} = 60.66\%$), movies ($\overline{F^m_{score}} = 80.8\%$) and news ($\overline{F^m_{score}} = 85\%$) which can be assigned to the specific rhythm and color diversity of each genre. However, these preliminary results show that in general each genre distinguishes itself from the others by a specific set of descriptors.

6 Conclusions and future work

We addressed the issue of automatic classification of video genres and proposed several types of content descriptors, namely: temporal, color and contour structural parameters. These descriptors are used to retrieve 7 common video genres (tests were performed on 91 hours of video footage in total).

At individual level, all genres achieve precision and recall ratios above 80%, while some genres achieve even higher detection ratios, close to 100%. Overall, for all genres, the combination of all descriptors, i.e. contour-color-action, provided the highest accuracy and achieves an average Fscore ratio above 80%, while the average correct detection ratio is above 94%. Finally, at feature level, average Fscore ratios are up to 92%.

One limitation of this approach is in it's computational complexity. Color and action descriptors rely mainly on temporal segmentation (cuts, fades, dissolves) which can be time consuming, while contours are extracted at different levels and re-refined after extracting the descriptors. Despite the fact that all processing, i.e. contour-color-action, takes less than half the sequence duration, to be integrated with a real indexing system, hardware acceleration/optimization is required.

However, these represent some of our preliminary work. Future work will include reducing data redundancy, addressing higher semantic levels of description (e.g. exploiting concept detection) as well as extending tests on a larger scale database.

Acknowledgments. The work has been co-funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Romanian Ministry of Labour, Family and Social Protection through the Financial Agreement POSDRU/89/1.5/S/62557.

References

1. A. F. Smeaton, P. Over, W. Kraaij, "High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements, Multimedia Content Analysis", Theory and Applications, Springer Verlag-Berlin, pp. 151-174, ISBN 978-0-387-76567-9, 2009.
2. D. Brezeale, D.J. Cook, "Automatic Video Classification: A Survey of the Literature", IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 38(3), pp. 416-430, 2008.
3. M.J. Roach, J.S.D. Mason, "Video Genre Classification using Dynamics", IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1557-1560, Utah, USA, 2001.
4. X. Yuan, W. Lai, T. Mei, X.S. Hua, X.Q. Wu, S. Li, "Automatic video genre categorization using hierarchical SVM", IEEE International Conference on Image Processing, pp. 2905-2908, Atlanta, USA, 2006.
5. M. Montagnuolo, A. Messina, "Parallel Neural Networks for Multimodal Video Genre Classification", Multimedia Tools and Applications, 41(1), pp. 125-159, 2009.
6. B. Ionescu, D. Coquin, P. Lambert, V. Buzuloiu: "A Fuzzy Color-Based Approach for Understanding Animated Movies Content in the Indexing Task", Eurasip Journal on Image and Video Processing, doi:10.1155/2008/849625, 2008.
7. C. Rasche: "An Approach to the Parameterization of Structure for Fast Categorization", International Journal of Computer Vision, 87(3), pp. 337-356, 2010.
8. B. Ionescu, A. Pacureanu, P. Lambert, C. Vertan, "Highlighting Action Content in Animated Movies", IEEE ISSCS - International Symposium on Signals, Circuits and Systems, 9-10 July, Iasi, Romania, 2009.
9. B. Ionescu, D. Coquin, P. Lambert, V. Buzuloiu, "Semantic Characterization of Animation Movies Based on Fuzzy Action and Color Information", AMR 4th International Workshop on Adaptive Multimedia Retrieval, Université de Genève, Switzerland, 2006.
10. B. Ionescu, V. Buzuloiu, P. Lambert, D. Coquin, "Improved Cut Detection for the Segmentation of Animation Movies", IEEE International Conference on Acoustic, Speech and Signal Processing, Toulouse, France, 2006.
11. W.A.C. Fernando, C.N. Canagarajah, D.R. Bull, "Fade and Dissolve Detection in Uncompressed and Compressed Video Sequence", IEEE International Conference on Image Processing, Kobe, Japan, pp. 299-303, 1999.
12. C.-W. Su, H.-Y.M. Liao, H.-R. Tyan, K.-C. Fan, L.-H. Chen, "A Motion-Tolerant Dissolve Detection Algorithm", IEEE Transactions on Multimedia, 7(6), pp. 1106-1113, 2005.
13. H.W. Chen, J.-H. Kuo, W.-T. Chu, J.-L. Wu, "Action Movies Segmentation and Summarization based on Tempo Analysis", ACM International Workshop on Multimedia Information Retrieval, pp. 251-258, New York, 2004.
14. R. W. Floyd and L. Steinberg, "An Adaptive Algorithm for Spatial Gray Scale", Proc. SID Int. Symp. Digest of Technical Papers, pp. 3637, 1975.
15. I. H. Witten, E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques", Second Edition, Eds. Morgan Kaufmann, ISBN 0-12-088407-0, 2005.
16. J. Canny, "A Computational Approach To Edge Detection", IEEE Transaction on Pattern Analysis and Machine Intelligence, 8(6), pp. 679698, 1986.