A color-action perceptual approach to the classification of animated movies

Bogdan Ionescu
LAPI - University Politehnica of Bucharest
061071 Bucharest, Romania
bionescu@alpha.imag.pub.ro

Constantin Vertan
LAPI - University Politehnica of Bucharest
061071 Bucharest, Romania
cvertan@alpha.imag.pub.ro

Patrick Lambert
LISTIC - Polytech Annecy-Chambery
B.P. 80439, 74944 France
patrick.lambert@univ-savoie.fr

Alexandre Benoit
LISTIC - Polytech Annecy-Chambery
B.P. 80439, 74944 France
alexandre.benoit@univ-savoie.fr

ABSTRACT

We address a particular case of video genre classification, namely the classification of animated movies. This task is achieved using two categories of content descriptors, temporal and color based, which are adapted to this particular content. Temporal descriptors, like rhythm or action, are quantifying the perception of the action content at different levels. Color descriptors are determined using color perception which is quantified in terms of statistics of color distribution, elementary hues, color properties (e.g. amount of light colors, cold colors, etc.) and color relationship. The potential of the proposed descriptors to the classification task has been proved through experimental tests conducted on more than 749 hours of video footage. Despite the high diversity of the video material, the proposed descriptors achieve an average precision and recall ratios up to 90% and 92%, respectively, and a global correct detection ratio up to 92%.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*color, action descriptors*; I.5.3 [**Pattern Recognition**]: Clustering—*video genre, animated movies*.

General Terms

Algorithms, Performances

Keywords

animated genre classification, action content, color properties, video indexing.

1. INTRODUCTION

Significant efforts are made to develop innovative automatic content-based indexing techniques to cope with challenges caused by accessing large collections of video footage. Of particular interest is the automatic cataloging of video footage into some predefined semantic categories. This can be performed *globally*, by classifying videos into one of several main genres, e.g. cartoons, music, news, sports. Also, sub-genres can be involved, e.g. identifying specific types of sports (football, hockey, etc.), movies (drama, thriller, etc.), and so on. Another solution aims at classifying movie content *locally*, thus considering video segments and specific concepts, e.g. outdoor vs. indoor, action, violence, etc. [1].

In this paper we address the global classification of a particular genre, namely the animated movies. The animated movie industry witnessed nowadays a spectacular development and gain in popularity: abundance of entertainment cartoon movies, festivals and expo, e.g. France - Annecy International Animated Film Festival, Canada - Ottawa International Animation Festival, Portugal - CINANIMA International Animation Film Festival, etc. Animated movies now target equally children and adults, becoming a distinctive industry similar to the artistic movies.

In the context of the automatic content-based retrieval, a common task related to this field is the automatic *selection of the "animated" content from other genres*. Regardless the approach, the main challenge is to derive attributes which are discriminant enough to distinguish between genres while maintaining a reduced dimensionality of the feature space. To this purpose, several approaches have been proposed in the literature.

One approach is to address the classification at *image level*. For instance, [2] emphasizes the basic characteristics of cartoons and uses nine color descriptors to distinguish between photographs and graphics over the World Wide Web. Another example is the approach in [3]. It uses Support Vector Machines (SVM) with several image descriptors, i.e. sat-

uration and brightness information, color histograms, edge information, compression ratio and pattern spectrum to label individual video frames as "cartoon" or "photographic". Authors announce correct classification ratios around 94% when tested on more than 24,000 static images. However, the main limitation of this approach is in its static nature, video specific dynamic information being disregarded.

Another category of approaches, which make the very subject of this paper, is to perform the classification at *sequence level*, e.g. [4] discusses an uni-modal approach and testes the prospective potential of motion information to cartoon classification. However, experimental validation was performed on a very limited data set, only 8 cartoon and 20 non cartoon sequences, making difficult to predict how the method will perform on a wider database. A two-modal approach is proposed in [5] and cartoon classification is performed using a multilayered perceptron with both visual (brightness, saturation, color hue, edge information, motion) and audio descriptors (MFCC descriptors). Tests were performed on a bit larger database containing 100 sequences (20 sequences of each genre: cartoons, commercials, music, news and sports) and classification accuracy is around 90%. Another example is the approach in [6] which uses eight human inspired MPEG-7 visual descriptors and a SVM scheme with active relevance feedback.

Other methods are addressing the *video genre classification*, which includes the case of cartoon movies. A state-of-the art is available in [9]. For instance, [10] proposes a truly multimodal approach which combines several types of content descriptors. 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 pieces of information are used for training a parallel neural network system and achieve a maximum accuracy rate up to 95% in distinguish between seven video genres (including cartoons): football, cartoons, music, weather forecast, newscast, talk shows and commercials. However, these techniques are not focusing on the retrieval of animated contents and are limited to use "all purpose" content descriptors which work with all genres but provide average performance with the animated content.

This short overview of the literature shows in general a lack of dedicated approaches, the few existing ones being limited to use more or less general purpose descriptors or a restraint testing framework. The remainder of this paper is organized as follows: Section 2 situates our work in the literature and highlights its contributions, Section 3 and Section 4 deal with feature extraction: temporal information and color properties. Experimental results are presented in Section 5 while Section 6 presents the conclusions and discuses future work.

2. THE PROPOSED APPROACH

The first limitation of the existing approaches is in the targeted genre which is exclusively the cartoon genre. In this paper we extend the classification by addressing, generically,

Figure 1: Various animation techniques (from left to right and top to bottom): paper drawing, object animation, 3D synthesis, glass painting, plasticine modeling and color salts (source CITIA [11]).

the animated movies and thus including equally cartoons and artistic animated movies. Artistic animated movies, less common than cartoons, but with an increasing popularity, are usually short animated clips having artistic connotations. Contrary to cartoons, artistic animated movies are produced using a high variety of techniques and use artistic concepts (see CITIA [11] and [7]). Some examples are depicted in Figure 1.

Due to their distinctive creation process and particular contents, animated movies often require a different processing approach than natural movies. Most of the existing approaches are proposing generic content descriptors, like image-based descriptors: e.g. saturation, brightness, contours or temporal-based: motion, sound, etc., which are not particularly addressing the properties of this genre. For instance, motion is usually discontinuous with animated movies and sometimes impossible to estimate, many artistic movies are without sound or commentaries, etc.

In this study we propose two categories of content descriptors, namely: *temporal-based* (animated movies usually have a different visual rhythm or action content) and *color-based* (color distribution is always specific), which are adapted to the animation content. For the *temporal descriptors*, e.g. rhythm, action, user experiments have been conducted on animated movies to quantify the perception of the action content at different levels. Temporal information is quantified in terms of visual rhythm, action content and amount of gradual transitions. On the other hand, the *color descriptors* have been validated on the semantic analysis of artistic animated movies [7] and are extended to temporal level (global descriptors). Using a color naming system, color perception is quantified 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.

In what concerns the evaluation, most of the existing approaches lack of a robust evaluation framework. Usually the test set is very limited and the evaluation measures are non standard or incomplete, e.g. only precision is announced, or accuracy, which in fact includes the correct detection of both classes, i.e animation and non animation, etc. To provide a pertinent evaluation we use an extensive data set,

namely 749 sequences containing various genres: animated movies, commercials, documentaries, movies, news broadcasting, sport and music. Classification is performed with several approaches: SVM (Support Vector Machines), KNN (K-Nearest Neighbor) and LDA (Linear Discriminant Analysis). Performance is assessed by computing average precision vs. recall curves, Fscore and correct detection ratios. In this "rough" evaluation framework, the proposed descriptors still achieve an average precision and recall ratios up to 90% and 92%, respectively, and a global correct detection ratio up to 92%. Content descriptors are described with the following sections.

3. ACTION DESCRIPTORS

The first feature set aims to capture the movie's temporal structure in terms of *visual rhythm*, *action content* and *amount of gradual video transitions*, as these parameters are strongly related to movie contents. To do so, first we perform a temporal segmentation, which roughly means parsing the movie intro shots by means of detecting the video transitions. We detect cuts and two of the most frequent gradual transitions, i.e. fades and dissolves. To favor the animated movies, we use specially adapted algorithms: cut detection is performed using the histogram-based approach proposed in [12], while fade and dissolve detection are carried out using an adaptation of the pixel-level statistical approach proposed in [14] and of the analysis of fading-in and fading-out pixels proposed in [15], respectively (more details are to be found at `http://imag.pub.ro/VideoIndexingRP2/`). Further, we determine the following parameters:

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, thus:

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

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:

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

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.

Defined in this way, $\bar{v}_T$ represents the average number of shot changes over the time interval T for the entire movie, being a measure of the movie global tempo. High values of $\bar{v}_T$ indicate a movie with a general high change ratio, while small values correspond typically to movies with predominant long and static shots (a reduced number of scenes).

Action. To determine the following parameters, we use a relatively confirmed 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 (denoted "hot action") and video segments with a low action content (the opposite situation).

Table 1: Movie rhythm versus action content.

Movie	Segment [frames]	Length [s]	$\bar{v}_T$
"Hot action"			
François le Vaillant	2961-3443	19	3.51
	9581-10134	22	3.82
	11456-11812	14	3.25
Ferrailles	5303-5444	6	5
	8391-8657	11	3.38
Circuit Marine	7113-7401	11	3.7
The Lyon and the Song	14981-15271	12	2.33
Toy Story	2917-3582	27	2.75
	99962-101090	45	3.84
	101710-102180	19	4.43
Le Moine et le Poisson	6428-6775	14	3.5
"Low action"			
Le Trop Petit Prince	633-1574	38	0.31
	6945-8091	46	0.37
François le Vaillant	4257-6523	91	0.18
	6898-7683	31	0.38
A Bug's Life	4662-5535	35	0.17
	37209-38769	62	0.62
	66027-67481	58	0.46

Table 2: Action groundtruth.

Action type	"hot action"	"low action"
$E\{\bar{v}_T\}$	3.65	0.48
$\sigma_{\bar{v}_T}$	0.85	0.23
interval	2.8-∞	0.25-0.71

We tune the method parameters in order to adapt to the animated content. We have conducted an experimental test on a small set of animated movies (8 movies from CITIA [11] and Pixar Animation Company). Ten people were asked to manually browse movie contents and identify, if possible, frame segments (described as intervals $[frame_A; frame_B]$) which best fits the two generic action categories, namely: "*hot action*" (corresponding to movie segments with an intense action content, e.g. fast changes, fast motion, visual effects, etc.) and "*low action*" (mainly static scenes). To avoid inter-annotator consistency and thus repeat annotation, each person annotated different video parts or sequences. For each manually labeled action segment, we compute the mean shot change ratio, $\bar{v}_T$ (see equation 1), to capture the corresponding changing rhythm. Some of the results are presented in Table 1. Then, we compute the overall $\bar{v}_T$ mean values over all the segments within each action category, as well as the standard deviation. Having these pieces of information, we determine the intervals of $\zeta_T(i)$ values which correspond to each type of action content, as $[E\{\bar{v}_T\} - \sigma_{\bar{v}_T}; E\{\bar{v}_T\} + \sigma_{\bar{v}_T}]$. The results are synthesized with Table 2.

Once we determine the correspondence between action perception and $\bar{v}_T$ values, we use the straightforward approach in [8] to highlight video segments which show a high number of shot changes, i.e. $\zeta_T > 2.8$ and thus candidates for "hot action" label, and a reduced number of shot changes, i.e. $\zeta_T < 0.71$ or corresponding to low action. To reduce over-segmentation of action segments, we merge neighboring action segments (within same label) 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, namely the hot-action ratio (denoted HA) and the low-action ratio (denoted LA), defined thus:

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

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. Gradual transitions have a well defined meaning in the movie's narration. For instance a dissolve may be used to change the time of the action, similarly, a fade is used to change the action or, used in a fade group, introduces a pause before changing the action place, etc. High amounts of gradual transitions are related to a specific movie contents, for instance many artistic animated movies basically replace cuts with gradual transitions, which confers mystery to the movie (see movies "Paradise", "Cœur de Secours", "Le Moine et le Poisson", [11]). Therefore, 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.

4. COLOR DESCRIPTORS

One of the main particularities of the animated content is in the color distribution. Contrary to natural movies, animated movies tend to have specific color palettes, highly saturated colors, a color distribution derived from the variation of very few hues, color contrasts, very uniform color regions, etc. Therefore, we aim at capturing these properties by describing the movie's global color contents such as using statistics of color distribution (e.g. cold, warm, saturated), elementary hues, color properties and relationship of colors. This is carried out using an adaptation of the approach proposed in [7].

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 sub-sequence 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 downsampled 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 [16]. The proposed color parameters are determined as follows.

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, $c_e \in \Gamma_e$ with Γ_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 color usage and second, color wealth by quantifying color variation and diversity. Using previously determined histogram information and the color naming dictionary (where colors are named according to the color's hue, saturation and intensity), 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"}.

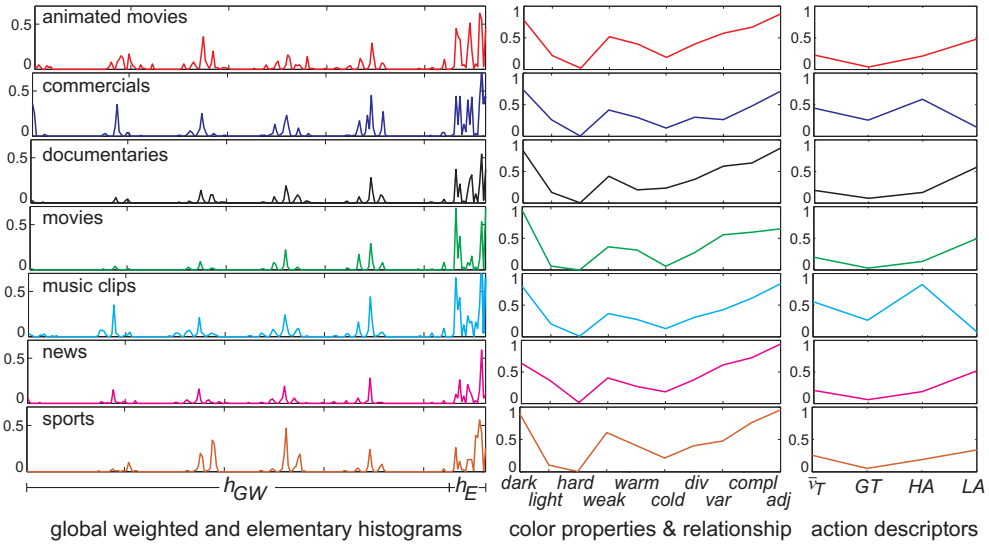

Figure 2: Average color-action feature vectors for each genre.

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:

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

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. 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), thus:

$$P_{adj} = \frac{Card\{c_e | Adj(c_e, c'_e) = True\}}{2 \cdot N_{ce}} \quad (9)$$

where $c_e \neq c'_e$ are the indexes of two significant elementary colors from the movie, $Adj()$ is the adjacency operator returning the true value if the two colors are analogous on Itten's color wheel, and N_{ce} is the movie's total number of elementary colors. Using the same reasoning, we define P_{compl} which reflects the amount of opposite perceptual color pairs (antipodal).

5. EXPERIMENTAL RESULTS

In order to obtain the most pertinent results, validation tests were conducted on a very large video database, i.e. 749 clips, with a high diversity of genres and sub-genres (more than 159 hours of video footage retrieved mainly from several TV chains).

The animated genre is represented with 209 sequences (54 hours) containing: artistic animated movies (source CITIA [11]), films and cartoon series (source Disney, Pixar, DreamWorks animation companies). The non animated genre is represented with 541 sequences (105 hours), namely: 320 commercials (4 hours, source 1980th TV commercials and David Lynch clips; some clips are containing both animated graphics and natural scenes); 74 documentaries (32 hours, both outdoor and indoor series, source BBC, IMAX, Discovery Channel); 57 movies (43 hours, both long movies and soap series, e.g. Friends, X-Files); 43 news broadcasting (19 hours, source TVR Romanian National Television Channel); 16 sports (4 hours, mainly soccer and outdoor extreme sports); 30 music clips (3 hours, source MTV Channel: dance, pop, techno music).

Table 3: KNN on all action and color descriptors.

train (%)	# train seq.	# train anim.	# test seq.	# test anim.	P (%)	R (%)	$\overline{TP}$	$\overline{FP}$ #	pub.	doc.	mov.	news	sp.	mus.	$\overline{FN}$
10	75	21	674	188	68	66	124	58	24	5	15	5	9	0	64
20	150	42	599	167	76	70	117	37	12	4	10	4	7	0	50
30	225	63	524	146	80	73	106	26	7	4	6	3	6	0	40
40	300	84	449	125	84	74	92	17	5	3	3	2	4	0	33
50	375	105	374	104	87	75	78	12	4	2	2	1	3	0	26
60	450	126	299	83	88	76	63	9	3	2	1	1	2	0	20
70	524	147	225	62	89	77	48	6	2	1	1	1	1	0	14
80	599	168	150	41	89	78	32	4	1	1	1	0	1	0	9
90	674	189	75	20	90	79	16	2	1	1	0	0	0	0	4

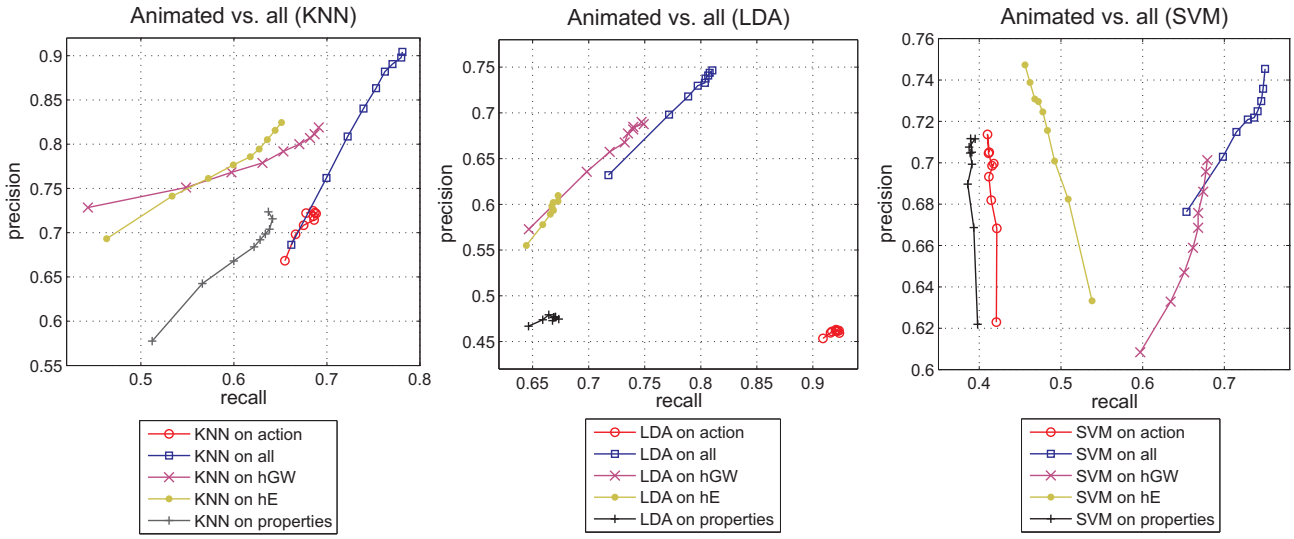

Figure 3: Precision vs. recall curves for different runs (action descriptors, h_{GW}, h_E, color properties and all parameters together) and amounts of training data (% of training is increasing along the curves).

5.1 Descriptor examples

To preliminary analyze the discriminant power of the proposed descriptors, Figure 2 depicts average color (see Section 4) and action (see Section 3) feature vectors for each genres. When compared to the other genres, the animated movies show a relatively different signature, e.g. have a different color pattern (more variations of basic hues being used, see the peaks in h_{GW}), most of the common hues are used in important amounts (see h_E), they tend to have a reduced global visual rhythm (see $\bar{v}_T$); while commercials and music clips have a high visual rhythm and action content (see $\bar{v}_T$ and HA), sports have a predominant hue (see the predominant peak in h_E), and so on. Discriminant power of the features is evidenced however in the classification task below.

5.2 Classification approach

Animated genre classification is carried out with a binary classification approach, i.e. considering two classes: animated and non animated. Each movie is represented with a feature vector, according to the previously presented content descriptors (several combinations are tested). For the classification we use three approaches, thus: the k-Nearest Neighbors algorithm (KNN, with k=5, cosine distance and majority rule), Support Vector Machines (SVM, with a linear kernel) and Linear Discriminant Analysis (LDA, applied on a PCA-reduced feature space) [17]. The method parameters were set to optimal values for this scenario after several preliminary tests.

As the choice of the training set may distort the accuracy of the results, we have adopted an exhaustive testing. Tests were performed for different amounts of training data (see the beginning of Table 3). For each set, tests are repeated using a cross validation approach, thus generating all possible combinations between training and test data, in order to shuffle all sequences.

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

$$P = \frac{\overline{TP}}{\overline{TP} + \overline{FP}}, \quad R = \frac{\overline{TP}}{\overline{TP} + \overline{FN}} \tag{10}$$

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, over all experimentations for a certain amount of training data (all combinations between test and training sequences).

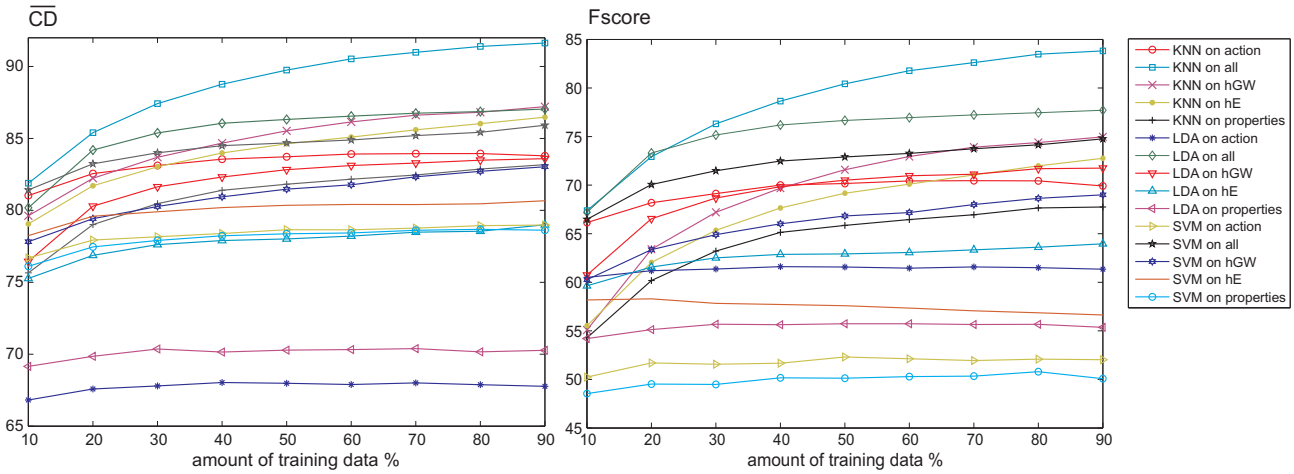

Figure 4: Average correct detection $\overline{CD}$ and F_{score} ratios for different amounts of training data.

On the other hand, to provide a global measure of performance we compute the average correct detection ratio, which takes into account good detections from both classes (animated and non animated), denoted $\overline{CD}$, and the F_{score} ratio which gives a measure of the overall false and non detections, thus:

$$\overline{CD} = \frac{\overline{N_{GD}}}{N_{total}}, \quad F_{score} = 2 \cdot \frac{P \cdot R}{P + R} \tag{11}$$

were $\overline{N_{GD}}$ is the average number of good classifications and $N_{total} = 749$ is the number of test sequences.

5.3 Precision and recall

Figure 3 depicts the obtained precision vs. recall curves for different amounts of training data and different runs. With KNN we obtain a precision and recall up to 90.1% and 78.6%, respectively (using all descriptors), with LDA the precision and recall are up to 74.7% (using all descriptors) and 92.4%, respectively (using only action descriptors) while with SVM precision and recall are up to 74.7% (using h_E histogram) and 74.9%, respectively (using all descriptors). Overall, the highest precision is achieved with KNN on all the descriptors, thus 90.1%, while the highest recall is obtained with LDA on action descriptors, namely 92.4%.

However, the best method in terms of both precision and recall proves to be KNN run on all action-color descriptors together. The resulted average precision, recall, $\overline{TP}$, $\overline{FP}$ and $\overline{FN}$ are presented in detail in Table 3 (for visualization purpose, actual real data values are to be rounded to nearest integer value). The results are very promising considering the diversity of video material (including a high variety of animated genres, see the beginning of Section 5) and also the size of test dataset.

For only 10% of training, average precision and recall are around 70% when testing on 674 sequences from which 188 are animated, while for 50% training precision approaches 90% and recall 80%. Also, one may observe the reduced number of false detections while maintaining a good detection ratio. For instance, using 70% training we obtain in average 48 good detections, only 6 false detections and 14 non detections.

To analyze which genres were wrongly classified in the animated category, we present in Table 3 the distribution of false positives to the other genres (we use the notations: pub. = commercials, doc. = documentaries, mov. = movies, sp. = sports, mus. = music).

From the six genres, the most distinctive proves to be the music genre. None of the clips were classified as animated (regardless the amount of training). This is due to their very distinctive color signature (typically darker colors due to the intensive use of visual effects) and a high visual rhythm (lot of changes over a short period of time). On the other hand, the most wrongly classified genre are the commercials. This is mainly because many of them involve a lot of computer graphics and animation (also there is a practical reason, the test database includes a lot of commercials, compared to the other genres). On the third place are the movies, which for a small amount of training tend to be confounded with animation (several movies are science fiction series, thus involving an abstract contents). These considerations are also predictable from the average signatures in Figure 2. Other genres, are misclassified occasionally.

Nevertheless, all false detections are dropping with the increase of the training set, being very reduced for an amount of training above 50%.

5.4 Global evaluation

Figure 4 depicts the obtained average correct detections ($\overline{CD}$) and the F_{score} ratios for different amounts of training and runs. Based on this information, the most powerful approach proves to be, again, the combination of all descriptors and KNN classification which is followed by LDA classification.

We obtained average $\overline{CD}$ and F_{score} ratios up to 91.63% and 83.82%, respectively. For only 50% of training data, correct detection ratio is above 90%, thus from 374 sequences more than 336 were labeled correctly in one of the two categories,

animated or non animated. The results are significant even for the lowest amount of training. For only 10% training data, i.e. 75 sequences (for all genres, see Table 3), from 764 test sequences 626 were correctly labeled into the two categories.

6. CONCLUSIONS AND FUTURE WORK

We addressed a particular case of video genre classification, i.e. the classification of the animated genre. We proposed two categories of content descriptors which are adapted to animated contents, namely: *temporal descriptors*, e.g. rhythm, action, for which user experiments have been conducted on animated movies to quantify the perception of the action content at different levels and *color descriptors* for which color perception is quantified in terms of statistics of color distribution, elementary hues, color properties (e.g. amount of light colors, cold colors, etc.) and color relationship.

These descriptors were used with several binary classification techniques to classify video footage into animated and non animated content. To provide a pertinent evaluation tests were performed on an extensive data set, namely 749 sequences containing various genres of animated movies, but also other video genres: commercials, documentaries, movies, news, sport and music. We achieve very promising results when using all descriptors together (considering the size of the test database and the diversity of video material) namely an average precision and recall ratios up to 90% and 92%, respectively, and a global correct detection ratio up to 92%.

However, these descriptors, alone, prove to be efficient for this particular classification task, being not discriminative enough to retrieve all other genres (through tests proved that genre classification requires a multimodal approach, e.g. using audio-visual features). Future work on this matter should push forward descriptors to a higher semantic level, like exploiting human concept detection.

7. ACKNOWLEDGMENTS

This 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. The authors would like to thank CITIA - The City of Moving Images and Folimage Animation Company for providing them with access to their animated movie database.

8. 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] V. Athitsos, M.J. Swain and C. Frankel, "Distinguishing photographs and graphics on the world wide web", IEEE Workshop on Content-Based Access of Image and Video Libraries, pp. 10-17, 1997.

[3] T.I. Ianeva, A.P. Vries and H. Rohrig, "Detecting cartoons: a case study in automatic video-genre classification", IEEE International Conference on Multimedia and Expo, 1, pp. 449-452, 2003.

[4] M. Roach, J. S. Mason, and M. Pawlewski, "Motion-based classication of cartoons", International Symposium on Intelligent Multimedia, pp. 146-149, 2001.

[5] R. Glasberg, A. Samour, K. Elazouzi and T. Sikora, "Cartoon-recognition using video and audio-descriptors", 13th European signal processing conference, Antalya, Turkey, 2005.

[6] X. Gao, J. Li and N. Zhang, "A Cartoon Video Detection Method Based on Active Relevance Feedback and SVM", Springer Lecture Notes in Computer Science, 3972, pp. 436-441, 2006.

[7] B. Ionescu, D. Coquin, P. Lambert and 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.

[8] B. Ionescu, L. Ott, P. Lambert, D. Coquin, A. Pacureanu and V. Buzuloiu, "Tackling Action - Based Video Abstraction of Animated Movies for Video Browsing", SPIE - Journal of Electronic Imaging, 19(3), 2010.

[9] 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.

[10] M. Montagnuolo, A. Messina, "Parallel Neural Networks for Multimodal Video Genre Classification", Multimedia Tools and Applications, 41(1), pp. 125-159, 2009.

[11] CITIA - Animaquid Animated Movie Indexing System, http://www.annecy.org/home/index.php?Page_ID=44.

[12] 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.

[13] H.W. Chen, J.-H. Kuo, W.-T. Chu and 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] 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.

[15] 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.

[16] R. W. Floyd and L. Steinberg, "An adaptive algorithm for spatial gray scale", Proceedings of Society for Information Display International Symposium, p. 3637, Washington, DC, USA, April 1975.

[17] I. H. Witten, E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques", Second Edition, Eds. Morgan Kaufmann, ISBN 0-12-088407-0, 2005.