**Noname manuscript No.**
(will be inserted by the editor)

# An Audio-Visual Approach to Web Video Categorization

**Bogdan Ionescu · Klaus Seyerlehner · Ionuţ Mironică · Constantin Vertan · Patrick Lambert**

Received: date / Accepted: date

**Abstract** In this paper, we discuss and audio-visual approach to automatic web video categorization. To this end, we propose content descriptors which exploit audio, temporal, and color content. The power of our descriptors was validated both in the context of a classification system and as part of an information retrieval approach. For this purpose, we used a real-world scenario, comprising 26 video categories from the blip.tv media platform (up to 421 hours of video footage). Additionally, to bridge the descriptor semantic gap, we propose a new relevance feedback technique which is based on hierarchical clustering. Experiments demonstrated that with this technique retrieval performance can be increased significantly and becomes comparable to that of high level semantic textual descriptors.

**Keywords** audio block-based descriptors · color perception · action assessment · video relevance feedback · video genre classification

B. Ionescu
LAPI, University "Politehnica" of Bucharest, 061071, Romania,
LISTIC, Polytech Annecy-Chambery, University of Savoie, 74944, France,
E-mail: bionescu@alpha.imag.pub.ro

K. Seyerlehner
DCP, Johannes Kepler University, A-4040 Austria,
E-mail: klaus.seyerlehner@jku.at

I. Mironică
LAPI, University "Politehnica" of Bucharest, 061071, Romania,
E-mail: imironica@alpha.imag.pub.ro

C. Vertan
LAPI, University "Politehnica" of Bucharest, 061071, Romania,
E-mail: constantin.vertan@upb.ro

P. Lambert
LISTIC, Polytech Annecy-Chambery, University of Savoie, 74944 France,
E-mail: patrick.lambert@univ-savoie.fr

## 1 Introduction

Automatic labeling of video footage according to genre is a common requirement in indexing large and heterogeneous collections of video material. This task can be tackled, either *globally* or *locally*. Global classification approaches aim to categorize videos into one of several main genres, such as cartoons, music, news, sports, documentaries, or with finer granularity into sub-genres, for instance, according to specific types of sports (e.g., football, hockey) or movie (e.g., drama, thriller). Local classification approaches, in contrast, label video segments instead of whole videos according to specific human-centered concepts, for instance, outdoor vs. indoor scenes, action segments, scenes showing violence (see TRECVid campaign [3]).

In this paper, we address the global classification task and consider the problem within a machine learning paradigm. In the literature, many sources of information have been exploited for this task [1]. A common approach is to use *text-based* information. Most existing web media search engines (e.g., YouTube, blip.tv) rely on text-based retrieval, as it provides a higher semantic level of description than other information sources. Text is obtained either from scene text (e.g., graphic text, sub-titles), from the transcripts of dialogues obtained with speech recognition techniques, or from other external sources, for instance, synopses, user tags, metadata. Common genre classification approaches include classic Bag-of-Words model [4] and Term Frequency-Inverse Document Frequency (TF-IDF) approaches [2].

Using audio-visual information is less accurate than using text. *Audio-based* information can be derived either from the time or from the frequency domain. Typical *time-domain* approaches include the use of Root Mean Square (RMS) of signal energy [25], sub-band information [5], Zero-Crossing Rate (ZCR) [27] or silence ratio. *Frequency-domain* features include energy distribution, frequency centroid [27], bandwidth, pitch [6] and Mel-Frequency Cepstral Coefficients (MFCC) [26].

The most popular type of audio-visual content descriptors are, however *visual descriptors*. They exploit both static and dynamic aspects of visual information either in the *spatial domain*, for instance, using color, temporal structure, objects, feature points, motion, or in the *compressed domain*, for example, using MPEG coefficients [1]. *Color descriptors* are generally derived at the image level and quantified via color histograms or other low-level parameters such as predominant color, color entropy, and variance (various color spaces are employed, e.g. RGB - Red Green Blue, HSV - Hue Saturation Value, and YCbCr - Luminance, Chrominance) [7] [8]. *Temporal structure-based descriptors* exploit temporal segmentation of video sequences. A video sequence is composed of several video shots connected by video transitions, which can be sharp (cuts) or gradual (fades, dissolves) [28]. Existing approaches basically exploit the frequency of their occurrence in the movie. Although some approaches use this information directly [9] (e.g., rhythm, average shot length), others derive features related to visual activity and exploit the concept of action (e.g., a high frequency of shot changes is often correlated with action) [21].

*Object-based features* in genre classification are generally limited to characterizing the occurrence of face and text regions in frames [9] [21]. Other related approaches exploit the presence of feature points, for example, using the well known SIFT descriptors [13]. *Motion-based descriptors* are derived either by motion detection techniques (foreground detection) or by motion estimation (i.e., prediction of pixel displacement vectors between frames). Typical features describe motion density, camera movement (global movement), or object trajectory [11]. Finally, less common are features computed in the *compressed video domain*, for example, using DCT (Discrete Cosine Transform) coefficients and embedded motion vectors from the MPEG flow [12]. Their main advantage is their availability with the initial video file.

All sources of information provide advantages and disadvantages. However, depending on the classification scenario, some prove to be more convenient than others. *Text-based* information, due to its high informational redundancy and reduced availability with visual information, can be less relevant when addressing a reduced number of genres (e.g., TV media genres). Also, it can produce high error rates if retrieved with speech transcription techniques [1]; however, it is the "golden standard" in web genre categorization; *object-based* information, although computationally expensive to process, tends to be semi-automatic (requires human confirmation); *motion information* tends to be available in high quantities during the entire sequence (object/camera), but is insufficient by itself to distinguish between specific genres, for instance, movies, sports, music [1]. *Audio-based* information provides good discriminative power for most common TV genres and requires fewer computational resources to be obtained and processed. *Color information* is not only simple to extract and inexpensive to process, but also very powerful in distinguishing cinematic principles and techniques; *temporal-based* information is a popular choice and proves to be powerful as long as efficient video transition detection algorithms are employed (e.g., adapting to web-specific low-quality video contents [22]).

The remainder of this paper is organized as follows: Section 2 discusses, and situates our work in relation to, several relevant genre classification approaches. Section 3 presents the proposed video descriptors (audio, temporal, and color-based). Section 4 discusses the improvement in retrieval performance achieved with relevance feedback, and proposes an approach inspired by hierarchical clustering. Experimental results are presented in Section 5, while Section 6 presents the conclusions and discusses future work.

## 2 Related work

Although, some sources of information provide better results than others in video genre categorization [1], the most reliable approaches - which also target a wider range of genres - are *multi-modal*, that is multi-source. In this section, we discuss the performance of several approaches we consider relevant for the present work - from single-modal (which are limited to coping with a reduced number of genres) to multi-modal (which target more complex

categorizations). We focus exclusively on approaches relying on audio-visual information - the subject of this study.

A simple, single-modal approach is that proposed in [14]. It addresses genre classification using only video dynamics, namely background camera motion and object motion. A single feature vector in the DCT-transformed space ensures low-pass filtering, orthogonality, and reduced feature dimension. A classifier based on a Gaussian Mixture Model (GMM) is then used to identify three common genres: sports, cartoons, and news. Despite the limited content information used, applying this approach to a reduced number of genres achieves detection errors below 6%. The authors of [21] used spatio-temporal information such as average shot length, cut percentage, average color difference, camera motion (temporal) and face frames ratio, average brightness, and color entropy (spatial). Genre classification is addressed at different levels according to a hierarchical ontology of video genres. Several classification schemes (Decision Trees and several SVM approaches) are used to classify video footage into the main genres movie, commercial, news, music, and sports, and further into sub-genres: movies into action, comedy, horror, and cartoons, and sports into baseball, football, volleyball, tennis, basketball, and soccer. The highest precision in video genre categorization is around 88.6% and in sub-genre categorization 97% for sports and up to 81.3% for movies.

However, truly multi-modal approaches also include audio information. For instance, the approach in [15] combines synchronized audio (14 Mel-Frequency Cepstral Coefficients - MFCC) and visual features (mean and standard deviation of motion vectors, MPEG-7 visual descriptors). Dimensionality of the feature vectors is reduced by means of Principal Component Analysis (PCA), and videos are classified with a GMM-based classifier. Tested with five common video genres, namely sports, cartoons, news, commercials, and music, this approach yields an average correct classification up to 86.5%. Another example is the approach proposed in [31]. Features are extracted from four sources: visual-perceptual information (color, texture, and motion), structural information (shot length, shot distribution, shot rhythm, shot clusters duration, and saturation), cognitive information (e.g., number, positions, and dimensions of faces) and aural information (transcribed text, sound characteristics). These features are used to train a parallel Neural Network, which achieves an accuracy of up to 95% in distinguishing between seven video genres and sub-genres, namely football, cartoons, music, weather forecast, newscast, talk shows, and commercials. A generic approach to video categorization was discussed in [16]. Each video document is modeled by a Temporal Relation Matrix (TRM) which describes the relationship between video segments, that is, temporal intervals related to the occurrence of a specific type of event. Events are defined based on the specificity of video features, such as speech, music, applause, speaker (audio) and color, texture, activity rate, face detection, costume (visual). TRMs provide a similarity measure between documents. Experimental tests with several classification approaches (mostly tree-based) and the six video genres news, soccer, TV series, documentary, TV games, and

movies resulted in individual genre $F_{score}$ ratios ranging from 40% to 100% (e.g., for a Random Forest with cross-validation).

In this paper, we propose three categories of content descriptors which exploit both audio and visual modalities. Although these sources of information have already been exploited, one of the novelties of our approach is the way we compute the descriptors. The proposed *audio features* are block-level-based and have the advantage of capturing local temporal information by analyzing sequences of consecutive frames in a time-frequency representation. *Visual information* is described with temporal information and color properties. Temporal descriptors are derived using a classic confirmed approach, that is, analysis of shot change frequency [21] [31]. Further, we introduce a novel way of assessing action content by considering human perception. We seek to capture aspects of color perception with our color descriptors. Instead of the typical low-level color descriptors (e.g., predominant color, color variance, color entropy and frame-based histograms [21]), we project histogram features onto a standard human color naming system and determine descriptors such as the percentage of light colors, cold colors, saturated colors, color contrasts, and elementary hue distribution. This achieves a higher semantic level of description. A preliminary validation of the proposed descriptors classifying seven common TV genres (i.e., animated movies, commercials, documentaries, movies, music videos, news broadcast, and sports) yielded average precision and recall ratios of $87\% - 100\%$ and $77\% - 100\%$, respectively [18].

We extended and adapted this approach to the categorization of web video genres. Several experimental tests conducted on a real-world scenario - using up to 26 genres provided by the blip.tv media platform and approximately 421 hours of video footage - demonstrated the power of our audio-visual descriptors in this classification task. Tests were conducted both in the context of a classification system and as part of an information retrieval approach. To bridge the semantic gap, we also investigated the potential use of user expertise and propose a new relevance feedback technique which is based on hierarchical clustering. This allows us to boost retrieval performance of the audio-visual descriptors close to that obtained with high-level semantic textual information.

## 3 Content description

As previously mentioned, we use both, audio and visual information to classify video genres. From the existing modalities we exploit the *audio soundtrack*, *temporal structure*, and *color content*.

Our selection is motivated by the specificity of these information sources with respect to video genre. For instance, most common video genres have very specific audio signatures: music clips contain music, there is a higher prevalence of monologues/dialogues in news broadcasts, documentaries have a mixture of natural sounds, speech, and ambient music, in sports there is crowd noise, and so on. Considered visually, temporal structure and colors highlight specific genre contents; for instance, commercials and music clips

tend to have a high visual tempo, music clips and movies tend to have darker colors (mainly due to the use of special effects), commercials use many gradual transitions, documentaries have reduced action content, animated movies have specific color palettes and color contrasts, sports usually have a predominant hue (e.g., green for soccer, white for ice hockey), in news broadcasting an anchor is present (high frequency of faces).

The proposed content descriptors are to be determined globally, thus covering the complete sequence. Each modality results in a feature vector. This approach has the advantage of facilitating data fusion by simple concatenation of the resulting data. Below we describe each category of descriptors and emphasize their advantages.

### 3.1 Audio descriptors

To address the range of video genres, we propose audio descriptors which are related to rhythm, timbre, onset strength, noisiness and vocal aspects [20]. The proposed set of audio descriptors, called *block-level audio features*, have the key advantage of capturing temporal information from the audio track at a local level. Standard spectral audio features, such as Mel Frequency Spectral Coefficient, Spectral Centroid, and Spectral Roll Off, are commonly extracted from each spectral frame of the time-frequency representation of an audio signal (capturing a time span of 20 ms). The features we propose are computed from sequences of consecutive spectral frames called *blocks*. Depending on the feature, a block consists of 10 to up to 512 consecutive spectral frames. Thus, local features can themselves capture temporal properties (e.g., rhythmic aspects) of an audio track over a time span ranging from half a second up to 12 seconds of audio.

Blocks are analyzed at a constant rate and their frames overlap by default by 50%. We determine one local feature vector per block. These local vectors are then summarized by computing simple statistics separately for each dimension of the local feature vectors (e.g., depending on the feature, we use mean, variance, or median). A schematic diagram of this procedure is depicted in Figure 1.

First, the audio track is converted into a $22kHz$ mono signal. To obtain a perceptual time-frequency representation of the video soundtrack, we then compute the short-time Fourier transform and map the frequency axis according to the logarithmic cent-scale. Because human frequency perception is logarithmic. The resulting time-frequency representation consists of 97 logarithmically spaced frequency bands. Further, we derive the following complex block-level audio features:

**- spectral pattern** (1 block = 10 frames, 0.9 percentile statistics): characterize the timbre of the soundtrack by modeling those frequency components that are simultaneously active. The dynamic aspect of the signal is retained by sorting each frequency band of a block along the time axis. The block width

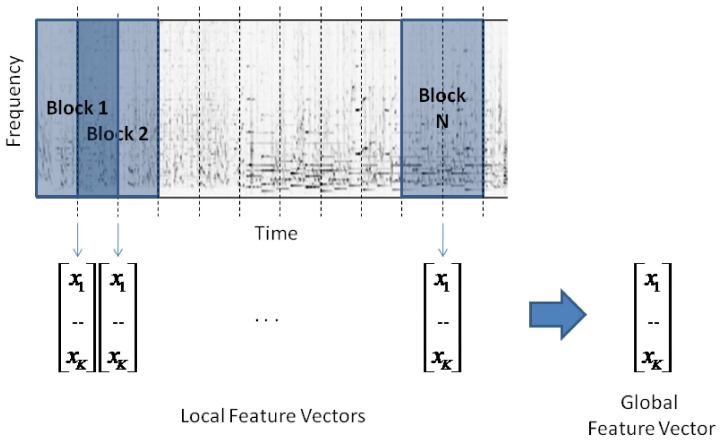


**Fig. 1** Processing a time ($OX$ axis) - frequency ($OY$ axis) representation in terms of spectral blocks ($N$ is the number of blocks).

varies depending on the extracted patterns, which allows capturing temporal information over different time spans.

**- delta spectral pattern** (1 block = 14 frames, 0.9 percentile statistics): captures the strength of onsets. To emphasize onsets, we first compute the difference between the original spectrum and a copy of the original spectrum delayed by 3 frames. As with the spectral pattern, each frequency band is then sorted along the time axis.

**- variance delta spectral pattern** (1 block = 14 frames, variance statistics): is basically an extension of the delta spectral pattern and captures the variation of the onset strength over time.

**- logarithmic fluctuation pattern** (1 block = 512 frames, 0.6 percentile statistics): captures the rhythmic aspects of the audio signal. In order to extract the amplitude modulations from the temporal envelope in each band, periodicities are detected by computing the Fast Fourier Transform (FFT) along each frequency band of a block. The periodicity dimension is then reduced from 256 to 37 logarithmically spaced periodicity bins.

**- spectral contrast pattern** (1 block = 40 frames, 0.1 percentile statistics): 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** (1 block = 256 frames, 0.5 percentile statistics). To capture the temporal relation of loudness changes over different frequency bands, we use the correlation coefficients between all possible pairs of frequency bands within a block. The resulting correlation matrix forms the correlation

pattern. The correlation coefficients are computed for a reduced frequency resolution of 52 bands.

These audio features in combination with a Support Vector Machine (SVM) classifier constitute a highly efficient automatic music classification system. At the 2010 Music Information Retrieval Evaluation eXchange, this approach ranked first in automatic music genre classification [20]. However, the proposed approach has not yet been applied to video genre classification.

### 3.2 Temporal structure descriptors

Temporal descriptors are derived using a classic confirmed approach, that is, analysis of the shot change frequency [21]. Unlike existing approaches, we refine the assessment of the action level on the basis of human perception.

One of the main factors contributing to the success of temporal descriptors is an accurate preceding temporal segmentation [28]. First, we detect both cuts and gradual transitions. Cuts are detected by means of an adaptation of the histogram-based approach proposed in [22]; fades and dissolves are detected using a pixel-level statistical approach [23] and the analysis of fading-in and fading-out pixels [24], respectively. Further, we compute the following descriptors:

**- rhythm**: capture the movie's tempo of visual change, we compute the relative number of shot changes occurring within a time interval of $T = 5s$, denoted $\zeta_T$. Then, the rhythm is defined as the movie average shot change ratio, $\bar{v}_T = E\{\zeta_T\}$.

**- action**: We aim to define two opposite situations: video segments with high action content (called "hot action", e.g., fast changes, fast motion, visual effects) with $\zeta_T > 3.1$, and video segments with low action content (i.e., containing mainly static scenes) with $\zeta_T < 0.6$. These thresholds were determined experimentally using user ground truth. A group of ten people was asked to manually browse the content of several TV movies and identify, if possible, frame segments (i.e., intervals $[frame_A; frame_B]$) which fall into the two action categories mentioned. To avoid inter-annotator consistency, each person annotated different video parts. For each manually labeled action segment, we computed the mean shot change ratio, $\bar{v}_T$, to capture the corresponding changing rhythm. Then we computed the average and standard deviation of $\bar{v}_T$ over all segments within each action category. Using this information as ground truth, we determine $\zeta_T$ intervals for each type of action content as $[E\{\bar{v}_T\} - \sigma_{\bar{v}_T}; E\{\bar{v}_T\} + \sigma_{\bar{v}_T}]$ and thus the two threshold limits (lower limit for high action and upper limit for low action).

Further, we quantify the action content using two parameters - hot-action ratio ($HA$) and low-action ratio ($LA$), determined by:

$$HA = \frac{T_{HA}}{T_{total}}, \quad LA = \frac{T_{LA}}{T_{total}} \tag{1}$$

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 transitions ratio**: Since high numbers of gradual transitions are generally related to a specific video content we compute:

$$GT = \frac{T_{dissolves} + T_{fade-in} + T_{fade-out}}{T_{total}} \tag{2}$$

where $T_X$ represents the total duration of all gradual transitions of type $X$. This provides information about editing techniques which are specific to a genre, such as movies or commercial clips.

### 3.3 Color descriptors

Color information is an important source for describing visual content. Most of the existing color-based genre classification approaches are limited to using intensity-based parameters or generic low-level color features such as average color differences, average brightness, average color entropy [21], variance of pixel intensity, standard deviation of gray level histograms, percentage of pixels with saturation above a certain threshold [29], lighting key [30], object color, and texture.

We propose a more sophisticated strategy which addresses the perception of color content [33]. A simple and efficient way to accomplish this is using color names; associating names with colors allows everyone to create a mental image of a given color or color mixture. We project colors onto a color naming system, and color properties are described using statistics of color distribution, elementary hue distribution, color visual properties (e.g., percentage of light colors, warm colors, saturated colors), and relationships between colors (adjacency and complementarity). Prior to parameter extraction, we use an error diffusion scheme to project colors onto a more manageable color palette - the non-dithering 216 color Webmaster palette (an efficient color naming system). Colors are represented by the following descriptors:

**- global weighted color histogram** is computed as the weighted sum of each shot color histogram:

$$h_{GW}(c) = \sum_{i=0}^{M} \left[ \frac{1}{N_i} \sum_{j=0}^{N_i} h^j_{shot_i}(c) \right] \cdot \frac{T_{shot_i}}{T_{total}} \tag{3}$$

where $M$ is the total number of video shots, $N_i$ is the total number of the retained frames for shot $i$ (we use temporal sub-sampling), $h^j_{shot_i}$ is the color histogram of frame $j$ from shot $i$, $c$ is a color index from the Webmaster palette (we use color reduction), and $T_{shot_i}$ is the length of shot $i$. The longer the shot, the more important its contribution to the global histogram of the movie.

**- elementary color histogram**: describes the distribution of elementary hues in the sequence:

$$h_E(c_e) = \sum_{c=0}^{215} h_{GW}(c)|_{Name(c_e) \subset Name(c)} \quad (4)$$

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**: We define several color ratios to describe color properties. For instance, the light color ratio, $P_{light}$, reflects the percentage of bright colors in the movie:

$$P_{light} = \sum_{c=0}^{215} h_{GW}(c)|_{W_{light} \subset Name(c)} \quad (5)$$

where $c$ is a color whose name contains one of the words defining brightness, and $W_{light} \in \{$"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 richness with two parameters: color variation, $P_{var}$, which is the number of significantly different colors, and color diversity, $P_{div}$, defined as the number of significantly different color hues [33].

**- color relationship**: we compute $P_{adj}$, the number of perceptually similar colors in the movie and $P_{compl}$, the number of perceptually opposite color pairs.

This level of description provides several advantages: the globally weighted color histogram, $h_{GW}$, extends the definition of static image histograms by taking into account the video temporal structure. Values describe percentages of colors appearing during the entire sequence, which provides a global color signature of the sequence. Further, with the elementary color histogram, $h_E$, we provide a projection of color to pure spectrum colors (hues), thus disregarding the saturation and intensity information. This mechanism ensures invariance to color fluctuations (e.g., illumination changes) and provides information about predominant hues.

Color property and color relationship ratios provide a more perceptual analysis of the color distribution by quantifying dark-light, warm-cold, saturated and perceptually similar (adjacent) - opposite (complementary) colors. Finally, color variability and diversity provide information on how much variability is in the color palette of a movie and its basic hues. For instance, the presence of many diverse colors may signify more vivid sequences.

## 4 Relevance feedback

Following this content description methodology, we investigated the potential use of Relevance Feedback (RF) techniques in bridging the inherent semantic

gap that results from the automatic nature of the annotation process. Globally, a typical RF scenario can be formulated thus: for a certain retrieval query, a user provides feedback by marking the results as *relevant or non-relevant*. Then, the system computes a better representation of the information needed based on this ground truth, and retrieval is further refined. This process can go through one or more such iterations [35].

In the literature, many approaches have been investigated. One of the earliest and most successful relevance feedback algorithms is the Rocchio algorithm [36]. It updates the query features by adjusting the position of the original query in the feature space according to the positive and negative examples and their associated importance factors. Another example is the Feature Relevance Estimation (FRE) approach [38], which assumes for a given query that a user may consider some specific features more important than others. Every feature is given an importance weight such that features with greater variance have lower importance than elements with smaller variations. More recently, machine learning techniques have been introduced to relevance feedback approaches. Some of the most successful techniques use Support Vector Machines [39], classification trees, such as Decision Trees [40], Random Forest [42] or boosting techniques, such as AdaBoost [41]. The relevance feedback problem can be formulated either as a two-class classification of the negative and positive samples or as a one-class classification problem (i.e., separating positive samples from negative samples).

We propose an RF approach that is based on Hierarchical Clustering (HC) [37]. A typical agglomerative HC strategy starts by assigning one cluster to each object in the feature space. Then, similar clusters are progressively merged based on the evaluation of a specified distance metric. By repeating this process, HC produces a dendrogram of the objects, which may be useful for displaying data and discovering data relationships. This clustering mechanism can be very valuable in solving the RF problem by providing a mechanism to refine the relevant and non-relevant clusters in the query results. A hierarchical representation of the similarity between objects in the two relevance classes allows us to select an optimal level from the dendrogram which provides a better separation of the two than the initial retrieval.

The proposed hierarchical clustering relevance feedback (HCRF) is based on the general assumption that the video content descriptors provide sufficient representative power that, within the first window of retrieved video sequences, there are at least some videos relevant to the query that can be used as positive feedback. This can be ensured by adjusting the size of the initial feedback window. Also, in most cases, there is at least one non-relevant video that can be used as negative feedback. The algorithm comprises three steps: *retrieval*, *training*, and *updating*.

**Retrieval**. We provide an initial retrieval using a nearest-neighbor strategy. We return a ranked list of the $N_{RV}$ videos most similar to the query video using the Euclidean distance between features. This constitutes the initial RF

window. Then, the user provides feedback by marking relevant results, which triggers the actual HCRF mechanism.

**Training**. The first step of the RF algorithm consists of initializing the clusters. At this point, each cluster contains a single video from the initial RF window. Basically, we attempt to create two dendrograms, one for relevant and one for non-relevant videos. For optimization reasons, we use a single global cluster similarity matrix for both dendrograms. To assess similarity, we compute the Euclidean distance between cluster centroids (which, compared to the use of min, max, and average distances, provided the best results). Once we have determined the initial cluster similarity matrix, we attempt to merge progressively clusters from the same relevance class (according to user feedback) using a minimum distance criterion. The process is repeated until the number of remaining clusters becomes relevant to the video categories in the retrieved window (regulated by a threshold $\tau$).

**Updating**. After finishing the training phase, we begin to classify the next videos as relevant or non-relevant with respect to the previous clusters. A given video is classified as relevant or not relevant if it is within the minimum centroid distance to a cluster in the relevant or non-relevant video dendrogram.

**Algorithm 1** Hierarchical Clustering Relevance Feedback.

$N_{clusters} \leftarrow N_{RV}$; $clusters \leftarrow \{C_1, C_2, ..., C_{N_{clusters}}\}$;
**for** $i = 1 \rightarrow N_{clusters}$ **do**
  **for** $j = i \rightarrow N_{clusters}$ **do**
    *compute* $sim[i][j]$;
    $sim[j][i] \leftarrow sim[i][j]$;
  **end for**
**end for**
**while** $(N_{clusters} \geq \tau)$ **do**
  $\{min_i, min_j\} = argmin_{i,j}|_{C_i, C_j \in \{same\ relevance\ class\}}(sim[i][j])$;
  $N_{clusters} \leftarrow N_{clusters} - 1$;
  $C_{min} = C_{min_i} \cup C_{min_j}$;
  **for** $i = 1 \rightarrow N_{clusters}$ **do**
    *compute* $sim[i][min]$;
  **end for**
**end while**
$TP \leftarrow 0$; $current\_video \leftarrow N_{RV} + 1$;
**while** $((TP \leq \tau_1) \parallel (current\_video < \tau_2))$ **do**
  **for** $i = 1 \rightarrow N_{clusters}$ **do**
    *compute* $sim[i][current\_video]$;
  **end for**
  **if** ($current\_video$ is classified as relevant) **then**
    $TP \leftarrow TP + 1$;
  **end if**
  $current\_video \leftarrow current\_video + 1$;
**end while**

The entire RF process can be repeated if needed (e.g., if retrieval performance is still low) by acquiring new relevance feedback information from the user. Algorithm 1 summarizes the steps involved. The following notations

were used: $N_{RV}$ is the number of sequences in the browsing window, $N_{clusters}$ is the number of clusters, $sim[i][j]$ denotes the distance between clusters $C_i$ and $C_j$ (i.e., centroid distance), $\tau$ represents the minimum number of clusters which triggers the end of the training phase (set to a quarter of the number of sequences in a browsing window), $\tau_1$ is the maximum number of searched videos from the database (set to a quarter of the total number of videos in the database), $\tau_2$ is the maximum number of videos that can be classified as positive (set to the size of the browsing window), $TP$ is the number of videos classified as relevant, and *current_video* is the index of the currently analyzed video.

The main advantages of the proposed HCRF approach are implementation simplicity and speed because it is computationally more efficient than other clustering techniques, such as SVMs [39] (which also motivated the choice of HC as clustering method). Further, unlike most RF algorithms (e.g., FRE [38] and Rocchio [36]), it does not modify the query or the similarity. The remaining retrieved videos are simply clustered according to class label. HC has previously been used in RF but implemented differently. For instance, [43] proposed the QCluster algorithm for image retrieval. It generates a multi-point query to create a hierarchy of clusters followed by use of a Bayesian classification function. In our approach, we simply exploit the dendrogram representation of the two relevance classes. Experimental results are presented in Section 5.3.

## 5 Experimental results

Validation of the proposed content descriptors was carried out in the context of the MediaEval 2011 (Benchmarking Initiative for Multimedia Evaluation) Video Genre Tagging Task [2]. It addresses automatic categorization of web video genres from the blip.tv media platform (see `http://blip.tv/`).

The test data set consisted of 2375 sequences (around 421 hours of video footage) labeled according to 26 video genre categories (the numbers in brackets are the numbers of available sequences): "*art*" (66), "*autos and vehicles*" (36), "*business*" (41), "*citizen journalism*" (92), "*comedy*" (35), "*conferences and other events*" (42), "*documentary*" (25), "*educational*" (111), "*food and drink*" (63), "*gaming*" (41), "*health*" (60), "*literature*" (83), "*movies and television*" (77), "*music and entertainment*" (54), "*personal or autobiographical*" (13), "*politics*" (597), "*religion*" (117), "*school and education*" (11), "*sports*" (117), "*technology*" (194), "*environment*" (33), "*mainstream media*" (47), "*travel*" (62), "*video blogging*" (70), "*web development and sites*" (40) and "*default category*" (248, comprises movies that cannot be assigned to any of the previous categories). For more details on genre categories see [2].

The main challenge in classifying these videos lies in the high number of different genres with which to cope. Also, each genre category has a high variety of video material, which makes training difficult. Finally, video content available on web video platforms is typically video reports, and differs from

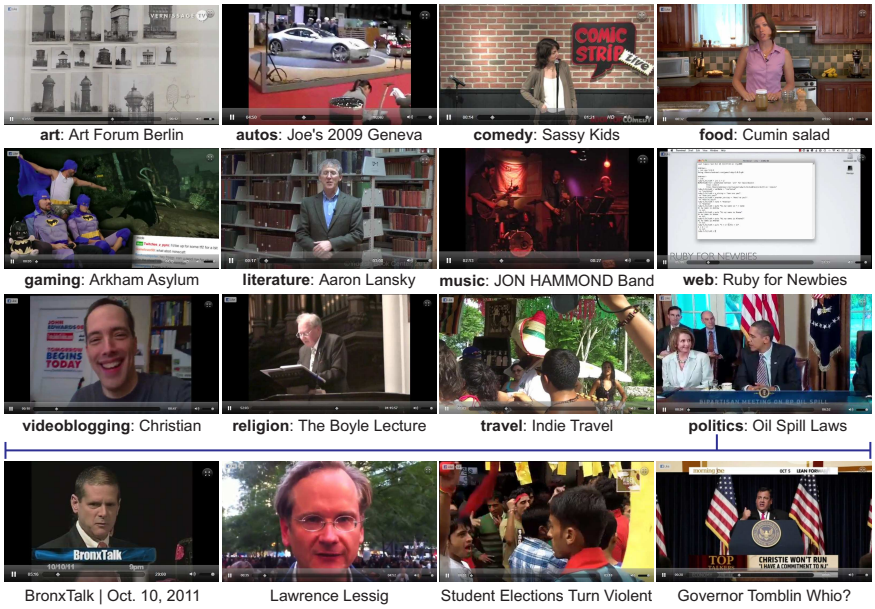


**Fig. 2** Image examples of several video genres. The bottom-row images exemplify the diversity of the "politics" category (source: blip.tv).

classic TV footage. Video material is usually assembled in a news broadcasting style, which means genre-specific content is inserted periodically into a dialogue or interview scene. Figure 2 illustrates these aspects.

Prior to video processing, we introduced a basic normalization step by converting all sequences to a reference video format. For genre categorization, each movie was represented by a feature vector which corresponds to the previously presented content descriptors. This yielded 9448 values for the audio descriptors and 245 for the color-action parameters, resulting in a total dimensionality of 9693 values. Data fusion was carried out using simple vector concatenation, which corresponds to early fusion approaches [34]. Below we describe each experiment in detail.

### 5.1 Classification perspective

**Experimental setup**. In the first experiment, we addressed video genre categorization in terms of machine learning techniques. We attempted to regroup the data according to genre-related clusters. For classification we used the Weka [19] environment, which provides a great selection of existing machine learning techniques. We tested methods ranging from simple Bayes to function-based, rule-based, lazy classifiers and tree approaches (from each category of methods, we selected the most representatives). Method parameters were tuned on the basis of preliminary experiments.

As the choice of training data may distort the accuracy of the results, we used a cross-validation approach. We split the data set into training and test sets, using values ranging from 10% to 90% for the percentage split. For part of the training data classification was repeated for all possible combinations between training and test sets in order to shuffle all sequences. Additionally, we tested different combinations of descriptors.

To assess performance, we used several measures. At the genre level, we computed average precision ($P$) and recall ($R$) (averaged over all experiments for a given percentage split), which account for the number of false classifications and misclassifications, respectively:

$$P = \frac{\overline{TP}}{\overline{TP} + \overline{FP}}, \quad R = \frac{\overline{TP}}{\overline{TP} + \overline{FN}} \tag{6}$$

where $\overline{TP}$, $\overline{FP}$, and $\overline{FN}$ represent the *average* numbers of true positives, false positives, and false negatives, respectively. As a global measure, we computed $F_{score}$ and average correct classification ($\overline{CD}$):

$$F_{score} = 2 \cdot \frac{P \cdot R}{P + R}, \quad \overline{CD} = \frac{\overline{N_{GD}}}{N_{total}} \tag{7}$$

where $\overline{N_{GD}}$ is the average number of correct classifications, and $N_{total}$ is the number of test sequences.

**Discussion of the results**. The most accurate classification was obtained by using all audio-visual descriptors in combination. For reasons of brevity, we present only these results. Figure 3 shows the overall average $F_{score}$ and average correct classification $\overline{CD}$ for a selection of seven machine learning techniques (those providing the most significant results).

The global results are very promising considering the high difficulty of this classification task. The highest average $F_{score}$ is 46.3%, while the best average correct classification is 55% (out of 475 test sequences 261 were correctly labeled, obtained for 80% training data). The most accurate classification technique proved to be an SVM with linear kernel, followed very closely by Functional Trees (FT), and then k-NN (with k=3), Random Forest trees, Radial Basis Function (RBF) Network, J48 decision tree, and finally Bayes Network (see Weka [19]).

The most interesting results, however, were obtained at genre level. Due to the high semantic content, not all genres can be classified correctly with audio-visual information. We sought to determine which categories are better suited for this approach. Figure 4 shows the genre average $F_{score}$ achieved by the linear SVM and FT trees.

The best performance was obtained for the following genres (we present results for a 50% percent split and give the highest value): "literature" ($F_{score}$ = 83%, highest 87%) and "politics" ($F_{score}$ = 81%, highest 84%), followed by "health" ($F_{score}$ = 78%, highest 85%), "citizen journalism" ($F_{score}$ = 65%, highest 68%), "food and drink" ($F_{score}$ = 62%, highest 77%), "web development and sites" ($F_{score}$ = 63%, highest 84%), "mainstream media"

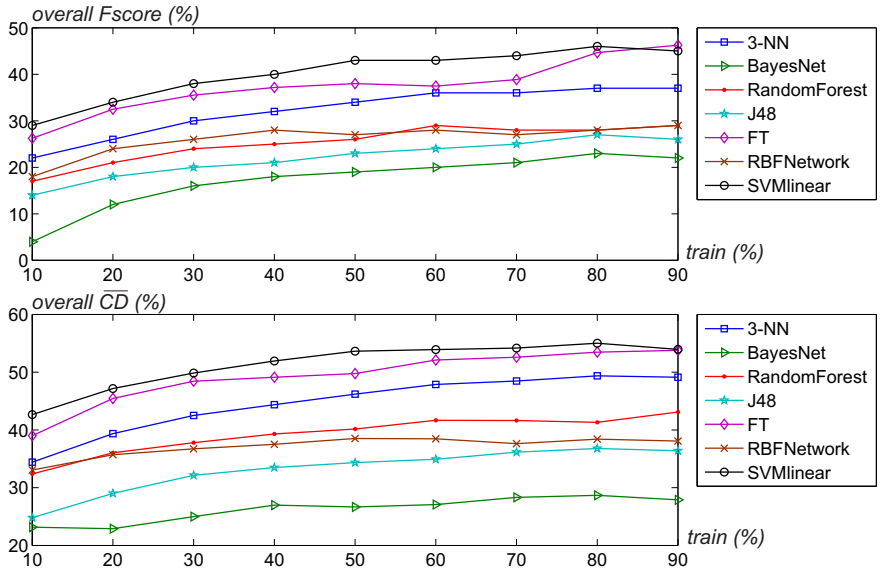


**Fig. 3** Overall average $F_{score}$ (see eq. 7) and overall average correct classification $\overline{CD}$ (see eq. 7) achieved by various machine learning techniques using all audio-visual descriptors.

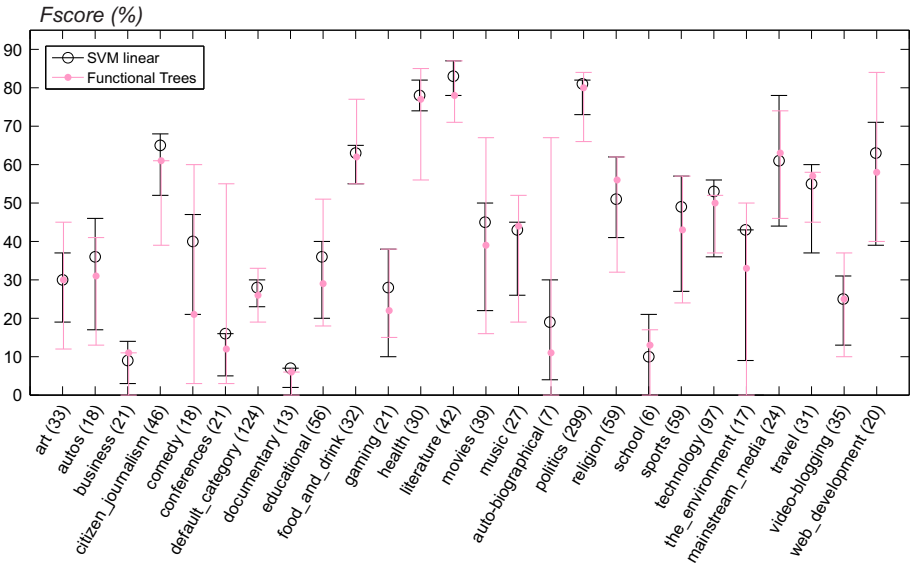


**Fig. 4** Average $F_{score}$ (see eq. 7) for linear SVM and Functional Trees (FT) using all audio-visual descriptors and for a training-test data percentage split of 50% (numbers in brackets are the numbers of test sequences in each genre). Vertical lines indicate the min-max $F_{score}$ intervals of each genre (percentage split ranging from 10% to 90%).

($F_{score} = 63\%$, highest 74%), "travel" ($F_{score} = 57\%$, highest 60%), "technology" ($F_{score} = 53\%$, highest 56%). Less successful performance was achieved for genres such as "documentary" ($F_{score} = 7\%$ which is also the highest), "school" ($F_{score} = 10\%$, highest 22%) or "business" ($F_{score} = 9\%$, highest 14%).

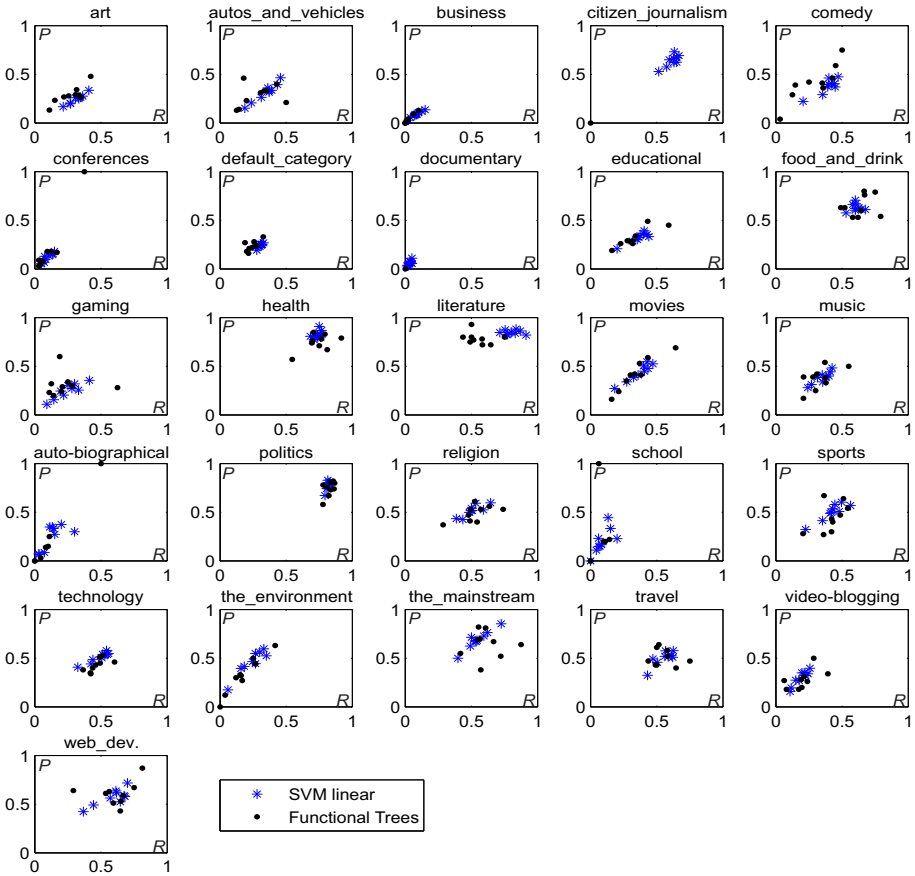


**Fig. 5** Average precision vs. recall (see eq. 6) achieved by SVM (linear kernel) and Functional Trees (FT) using all audio-visual descriptors and for various amounts of training data (percentage split from 10% to 90%).

Globally, classification performance increases with the amount of training data. However, for some genres, due to the large variety of video materials (see Figure 2), increasing the number of examples may result in overtraining and thus in reduced classification performance. It can be seen in Figure 3 that classification performance decreases as the proportion of training data increases (e.g., SVM linear for 90% training data). A clear difference between FT and SVM is visible at genre level. Globally, the SVM tends to perform

better on a reduced training set, while the FT tends to be superior for higher amounts of training data (e.g., training data $> 70\%$, see min-max intervals in Figure 4). This can also be observed for genre precision and recall.

Figure 5 shows genre average precision against recall for various percentage splits (ranging from 10% to 90%). The highest average precision, and thus the lowest number of false classifications, was achieved for the genres "literature" ($P = 93\%$ with FT), "health" ($P = 90.9\%$ with FT), "web development and sites" ($P = 87\%$ with FT), "the mainstream media" ($P = 85.3\%$ with SVM), "politics" ($P = 82.9\%$ with SVM), "food and drink" ($P = 79\%$ with FT), "comedy" ($P = 75\%$ with FT), "citizen journalism" ($P = 73\%$ with SVM), "movies and television" ($P = 69\%$ with FT) and "sports" ($P = 67\%$ with FT). The highest average recall was obtained for "literature" ($R = 91.3\%$ with SVM), "politics" ($R = 86.6\%$ with FT), "the mainstream media" ($R = 87.5\%$ with FT), "web development and sites" ($R = 81.3\%$ with FT), "food and drink" ($R = 79.2\%$ with FT) and "travel" ($R = 75\%$ with FT). Note that most of these values were obtained for the highest amount of training data (i.e., 90%).

### 5.2 Retrieval perspective

**Experimental setup**. In this experiment, we assessed the classification performance of the proposed descriptors in terms of an information retrieval system. We present the results obtained for the MediaEval 2011 Video Genre Tagging Task [2]. The challenge was to develop a retrieval mechanism that works with all 26 genre categories. Each participant was provided with a development set consisting of 247 sequences, unequally distributed with respect to genre. Some genre categories were represented with very few (even just one or two) examples. This initial set was to serve as a reference point for developing the proposed solution. The participants were encouraged to build their own training sets if required by their approach. Consequently, to provide a consistent training data set for the classification task, we extended the data set to up to 648 sequences. Additional videos were retrieved from the same source (blip.tv), using genre-related keywords (we checked for duplicates in the official development and test sets). The final retrieval task was performed on a test set consisting of 1727 sequences. In this case, the training-classification steps are to be performed only once. Up to 10 teams competed at this task, each one submitting up to 5 different runs (3 were restricted to using only textual descriptors extracted from speech transcripts, user tags, and metadata). A detailed overview of the results was presented in [2].

In our case, the retrieval results were obtained using a binary ranking in which the maximum relevance of 1 is associated with the genre category into which the document was classified, while other genres have 0 relevance. To assess performance, we used the overall Mean Average Precision (MAP) as defined by TRECVid [3] (also see trec_eval scoring tool at `http://trec.`

nist.gov/trec_eval/):

$$MAP = \frac{1}{|Q|} \cdot \sum_{j=1}^{|Q|} \frac{1}{m_j} \cdot \sum_{k=1}^{m_j} P(R_{j,k}) \tag{8}$$

where $Q = \{q_1, ..., q_{|Q|}\}$ denotes a set of queries $q_j$ which are represented in the data set by $\{d_1, ..., d_{m_j}\}$ relevant documents, $R_{j_k}$ is the set of ranked retrieval results from the top result to document $d_k$, and $P()$ is the precision (see eq. 6). When a relevant document is not retrieved at all, the precision value in the above equation is taken to be 0.

**Discussion of the results**. For classification we used the approach providing the most accurate results, namely the SVM with a linear kernel. In Table 1 we compare our results with several other approaches using various modalities of the video, from textual information (e.g., speech transcripts, user tags, metadata) to audio-visual[1].

Our descriptors achieved an overall MAP of up to 12% (see team RAF in Table 1). These were the best results obtained using audio-visual information alone. Use of descriptors such as cognitive information (face statistics), temporal information (average shot duration, distribution of shot lengths) [31], audio (MFCC, zero-crossing rate, signal energy), color (histograms, color moments, autocorrelogram - denoted autocorr.), and texture (co-occurrence - denoted co-occ., wavelet texture grid, edge histograms) with SVM resulted in a MAP of less than 1% (see team KIT in Table 1), while clustered SURF features and SVM achieved a MAP of up to 9.4% (see team TUB in Table 1). We achieved better performance even compared to some classic text-based approaches, such as the Term Frequency-Inverse Document Frequency (TF-IDF - MAP 9.8%, see team UAB in Table 1) and the Bag-of-Words (MAP 5.5%, see team SINAI in Table 1) approaches. Compared to visual information, audio descriptors seem to provide better discriminative power for this task.

It must be noted that the results presented in Table 1 cannot be definitive, as the classification approaches were not trained and set up strictly comparably. Teams were allowed to access other sources of information than those proposed in the competition. For instance, we used 648 sequences for training, whereas team KIT used up to 2514 sequences. Most text-based approaches employed query expansion techniques (e.g., Wordnet - see http://wordnet.princeton.edu/ and Wikipedia - see http://en.wikipedia.org). However, these results not only provide a (crude) performance ranking, but also illustrate the difficulty of this task.

The most efficient retrieval approach remains the inclusion of textual information, as it provides a higher semantic level of description than audio-visual information. The average MAP achieved by including textual descriptors is around 30% (e.g., see team TUB in Table 1). Retrieval performance is boosted by including information such as movie names, movie ID from *blip.tv*, or the

[1] the following notations were adopted: Terrier IR is an information retrieval system, see http://terrier.org/; Delicious is a social tagging site, see http://del.icio.us/.

**Table 1** Comparative results: MediaEval benchmarking [2] (selective results).

| descriptors | modality | method | decision | MAP | team |
|---|---|---|---|---|---|
| speech transcripts | text | Support Vector Machines | ranked list | 11.79% | LIA |
| speech transcripts, metadata, user tags | text | Bag-of-Words + Terrier IR | ranked list | 11.15% | SINAI |
| speech transcripts | text | Bag-of-Words | ranked list | 5.47% | SINAI |
| speech transcripts, metadata, user tags | text | TF-IDF + cosine dist. | binary | 9.4% | UAB |
| speech transcripts, Delicious tags, metadata | text | BM25F [32] + Kullback - Leibler diverg. | ranked list | 11.11% | UNED |
| metadata | text | Negative multinomial diverg. | ranked list | 39.37% | TUD |
| MFCC, zero cross. rate, signal energy | audio | multiple SVMs | binary | 0.1% | KIT |
| **proposed** | audio | SVM with linear kernel | binary | 10.29% | RAF |
| clustered SURF | visual | Visual-Words + SVM with RBF kernel | binary | 9.43% | TUB |
| hist., moments, autocorr., co-occ., wavelet, edge hist. | visual | multiple SVMs | binary | 0.35% | KIT |
| cognitive (face statistics [31]) | visual | multiple SVMs | binary | 0.1% | KIT |
| structural (shot statistics [31]) | visual | multiple SVMs | binary | 0.3% | KIT |
| **proposed** | visual | SVM with linear kernel | binary | 3.84% | RAF |
| color, texture, aural, cognitive, structural | audio, visual | multiple SVMs | binary | 0.23% | KIT |
| **proposed** | audio, visual | SVM with linear kernel | binary | 12.08% | RAF |
| clustered SURF, metadata | visual, text | Naive Bayes, SVM + serial fusion | binary | 30.33% | TUB |

username of the video uploader; in this particular case, the reported MAP was up to 56% (which is also the highest obtained).

In the following experiment we sought to prove that, notwithstanding the superiority of text descriptors, audio-visual information also has great potential in classification tasks, but may benefit from additional help. To this end, we investigated the use of relevance feedback.

### 5.3 Relevance feedback perspective

**Experimental setup**. In the final experiment, we attempted to enhance retrieval by employing the proposed relevance feedback scheme as described in Section 4. For tests, we used the entire data set: all 2375 sequences. Each sequence was represented by the proposed audio-visual descriptors. The user feedback was simulated automatically from the known class membership of each video (i.e., the genre labels). Compared to real user feedback, this has the advantage of providing a fast and extensive simulation framework, which otherwise could not be achieved due to physical constraints (e.g., availability of a significant number of users) and inherent human errors (e.g., indecision, misperception). We use only one feedback session. Tests were conducted for various sizes of the user browsing window.

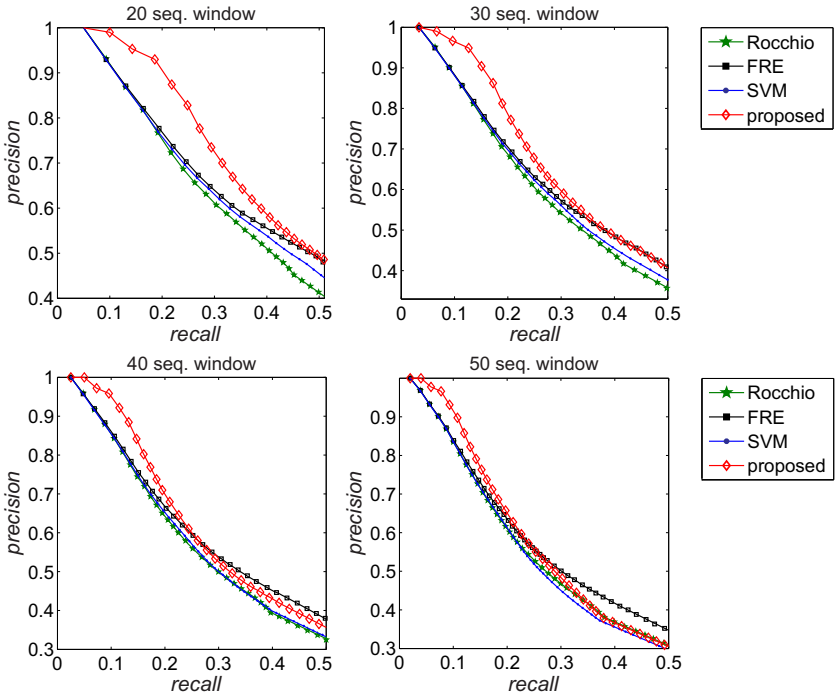


**Fig. 6** Precision - recall curves obtained with relevance feedback for different size browsing windows (for visualization purposes, we limited $OX$ to 0.5).

**Discussion of the results**. Figure 6 compares the precision - recall curves obtained with the proposed approach, hierarchical clustering relevance feedback (HCRF, see Section 4), with those of several other approaches, namely Rocchio [36], Feature Relevance Estimation (FRE) [38] and Support Vector Machines [39]. The proposed HCRF provides an improvement in retrieval, par-

ticularly for small browsing windows (e.g., 20, 30 video sequences, see the red line in Figure 6). With increasing window size, all methods tend to converge at some point to similar results.

Table 2 summarizes the overall retrieval MAP (see also eq. 8) estimated as the area under the uninterpolated precision-recall curve. For the proposed HCRF, the MAP ranges from 41.8% to 51.3%, which is an improvement over the other methods of at least a few percents. Also, it can be seen that relevance feedback proves to be a promising alternative for improving retrieval performance since it provides results close to those obtained with high-level textual descriptors (see Table 1).

**Table 2** MAP obtained with Relevance Feedback

| RF method | 20 seq. window | 30 seq. window | 40 seq. window | 50 seq. window |
|---|---|---|---|---|
| Rocchio | 46.8% | 43.84% | 42.05% | 40.73% |
| FRE | 48.45% | 45.27% | 43.67% | 42.12% |
| SVM | 47.73% | 44.44% | 42.17% | 40.26% |
| **proposed** | **51.27**% | **46.79**% | **43.96**% | **41.84**% |

## 6 Conclusions

We have addressed web video categorization using audio-visual information. We have proposed content descriptors which exploit audio, temporal structure, and color content and tested their power in solving this task. Experimental validation was carried out in the context of the MediaEval 2011 Video Genre Tagging Task [2], which addressed a real-world scenario - categorization of up to 26 video genres from the blip.tv media platform (421 hours of video footage). The tests were conducted both in the context of a classification system and as part of an information retrieval approach.

On classification, not all genres can be retrieved using audio-visual information. The use of audio-visual information may be highly efficient in detecting particular genres, for instance, in our case "literature" (we obtain $F_{score} = 87\%$), "politics" ($F_{score} = 84\%$), and "health" ($F_{score} = 85\%$), and less successful for others, such as "school" ($F_{score} = 22\%$), and "business" ($F_{score} = 14\%$). One can envisage a classification system which adapts the choice of parameters to the target categories, for instance, using audio-visual descriptors for genres which are best detected with this information, using text for text-related categories, and so on.

In retrieval, the proposed descriptors achieved the best results of all audio-visual descriptors. They provided better retrieval performance than other descriptors such as cognitive information (face statistics), temporal information

(average shot duration, distribution of shot lengths), audio (MFCC, zero-crossing rate, signal energy), color (histograms, color moments, autocorrelogram) and texture (co-occurrence, wavelet texture grid, edge histograms), and excelled even compared to some classic text-based approaches, such as the Term Frequency-Inverse Document Frequency (TF-IDF) approach. The results are, however, not definitive because different training data and different classifier setups were used. Text-based descriptors still achieve the best results. To address this, we designed a relevance feedback approach which allows boosting the performance close to that obtained with high-level semantic textual information (in this particular case, we achieve a MAP of up to 51%).

The main limitation of this approach, which is common to all ad-hoc genre categorization approaches, lies in the detection of genre related content. The proposed categorization system is limited to detect genre-related patterns, globally, such as identifying episodes from a series. It is not capable of detecting genre related segments within same sequence. Therefore, to provide good classification performance, each type of video material must be represented properly by the training set.

Future improvements will consist mainly of addressing sub-genre categorization and considering the constraints of very large scale approaches (millions of sequences and dozens of genre concepts). Also, we consider investigating the benefits to using relevance feedback with text-based retrieval.

## 7 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. We also acknowledge the 2011 Genre Tagging Task of the MediaEval Multimedia Benchmark [2] for providing the test data set.

## References

1. D. Brezeale, D.J. Cook, "Automatic Video Classification: A Survey of the Literature," IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 38(3), pp. 416-430, 2008.
2. M. Larson, A. Rae, C.-H. Demarty, C. Kofler, F. Metze, R. Troncy, V. Mezaris, Gareth J.F. Jones (eds.), Working Notes Proceedings of the MediaEval 2011 Workshop at Interspeech 2011, vol. 807, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-807/, Pisa, Italy, September 1-2, 2011.
3. 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.
4. G. Forman, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification," Journal of Machine Learning Research, 3, pp. 1289-1305, 2003.
5. M. H. Lee, S. Nepal, U. Srinivasan, "Edge-based Semantic Classification of Sports Video Sequences," IEEE Int. Conf. on Multimedia and Expo, 2, pp. 157-160, 2003.

6. J. Fan, H. Luo, J. Xiao, L. Wu, "Semantic Video Classification and Feature Subset Selection under Context and Concept Uncertainty," ACM/IEEE Conference on Digital Libraries, pp. 192-201, 2004.
7. X. Gibert, H. Li, D. Doermann, "Sports Video Classification using HMMs," Int. Conf. on Multimedia and Expo, 2, pp. II-345-348, 2003.
8. M. Ivanovici, N. Richard, "The Colour Fractal Dimension of Colour Fractal Images", IEEE Transactions on Image Processing, 20(1), pp. 227 - 235, 2010.
9. G. Wei, L. Agnihotri, N. Dimitrova, "TV Program Classification based on Face and Text Processing," IEEE Int. Conf. on Multimedia and Expo, 3, pp. 1345-1348, 2000.
10. X. Yuan, W. Lai, T. Mei, X.-S. Hua, X.-Q. Wu, S. Li, Automatic Video Genre Categorization using Hierarchical SVM," IEEE Int. Conf. on Image Processing, pp. 2905-2908, 2006.
11. G. Y. Hong, B. Fong, A. Fong, "An Intelligent Video Categorization Engine," Kybernetes, 34(6), pp. 784-802, 2005.
12. H. Wang, A. Divakaran, A. Vetro, S.-F. Chang, H. Sun, "Survey of Compressed-Domain Features used in Audio-Visual Indexing and Analysis," Journal of Visual Communication and Image Representation, 14(2), pp. 150-183, 2003.
13. Z. Wang, M. Zhao, Y. Song, S. Kumar, B. Li, "YouTubeCat: Learning to Categorize Wild Web Videos", In Proc. of Computer Vision and Pattern Recognition, pp. 879886, 2010.
14. M.J. Roach, J.S.D. Mason, "Video Genre Classification using Dynamics," IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 1557-1560, Utah, USA, 2001.
15. L.-Q. Xu, Y. Li, "Video classification using spatial-temporal features and PCA," International Conference on Multimedia and Expo, pp. 485488, 2003.
16. Z. Al A. Ibrahim, I. Ferrane, P. Joly, "A Similarity-Based Approach for Audiovisual Document Classification Using Temporal Relation Analysis," EURASIP Journal on Image and Video Processing, *doi* : 10.1155/2011/537372, 2011.
17. T. Sikora, "The MPEG-7 Visual Standard for Content Description An Overview," IEEE Trans. on Circuits and Systems for Video Technology, 11(6), pp. 696-702, 2001.
18. B. Ionescu, K. Seyerlehner, C. Rasche, C. Vertan, P. Lambert, "Content-based Video Description for Automatic Video Genre Categorization", The 18th International Conference on MultiMedia Modeling, 4-6 January, Klagenfurt, Austria, 2012.
19. Weka Data Mining with Open Source Machine Learning Software in Java, University of Waikato, `http://www.cs.waikato.ac.nz/ml/weka/`, 2011.
20. K. Seyerlehner, M. Schedl, T. Pohle, P. Knees, "Using Block-Level Features for Genre Classification, Tag Classification and Music Similarity Estimation," 6th Annual Music Information Retrieval Evaluation eXchange (MIREX-10), Utrecht, Netherlands, August 9-13, 2010.
21. X. Yuan, W. Lai, T. Mei, X.-S. Hua, X.-Q. Wu, S. Li, "Automatic Video Genre Categorization using Hierarchical SVM," IEEE Int. Conf. on Image Processing, pp. 2905-2908, 2006.
22. P. Kelm, S. Schmiedeke, T. Sikora, "Feature-based video key frame extraction for low quality video sequences", 10th Workshop on Image Analysis for Multimedia Interactive Services, pp.25-28, 6-8 May, London, UK, 2009.
23. W.A.C. Fernando, C.N. Canagarajah, D.R. Bull, "Fade and Dissolve Detection in Uncompressed and Compressed Video Sequence," IEEE Int. Conf. on Image Processing, Kobe, Japan, pp. 299-303, 1999.
24. C.-W. Su, H.-Y.M. Liao, H.-R. Tyan, K.-C. Fan, L.-H. Chen, "A Motion-Tolerant Dissolve Detection Algorithm," IEEE Trans. on Multimedia, 7(6), pp. 1106-1113, 2005.
25. Z. Rasheed, M. Shah, "Movie Genre Classification by Exploiting Audio-Visual Features of Previews," IEEE Int. Conf. on Pattern Recognition, 2, pp. 1086-1089, 2002.
26. U. Srinivasan, S. Pfeiffer, S. Nepal, M. Lee, L. Gu, S. Barrass, "A Survey of Mpeg-1 Audio, Video and Semantic Analysis Techniques," Multimedia Tools and Applications, 27(1), pp. 105-141, 2005.
27. Z. Liu, J. Huang, Y. Wang, "Classification of TV Programs based on Audio Information using Hidden Markov Model," IEEE Workshop on Multimedia Signal Processing, pp. 27-32, 1998.
28. R. Lienhart, "Reliable Transition Detection in Videos: A Survey and Practitioners Guide", Int. Journal of Image and Graphics, 1(3), pp. 469-486, 2001.

29. B. T. Truong, C. Dorai, S. Venkatesh, "Automatic Genre Identification for Content-Based Video Categorization," Int. Conf. on Pattern Recognition, IV, pp. 230-233, 2000.
30. Z. Rasheed, Y. Sheikh, M. Shah, "On the use of Computable Features for Film Classification," IEEE Trans. Circuits and Systems for Video Technology, 15, pp. 5264, 2003.
31. M. Montagnuolo, A. Messina, "Parallel Neural Networks for Multimodal Video Genre Classification", Multimedia Tools and Applications, 41(1), pp. 125-159, 2009.
32. J. Pérez-Iglesias, J. R. Pérez-Agüera, V. Fresno, Y. Z. Feinstein, "Integrating the Probabilistic Models BM25/BM25F into Lucene". CoRR, abs/0911.5046, 2009.
33. 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.
34. Cees G. M. Snoek, M. Worring, Arnold W. M. Smeulders, "Early versus Late Fusion in Semantic Video Analysis", ACM Int. Conf. on Multimedia, New York, USA, 2005.
35. C.D. Manning, P. Raghavan, H. Schütze, "Introduction to Information Retrieval", Cambridge University Press. 2008.
36. N. V. Nguyen, J.-M. Ogier, S. Tabbone, and A. Boucher, "Text Retrieval Relevance Feedback Techniques for Bag-of-Words Model in CBIR", International Conference on Machine Learning and Pattern Recognition, 2009.
37. W. J. Krzanowski, "Principles of Multivariate Analysis: A User's Perspective", Clarendon Press, Oxford, 1993.
38. Yong Rui, T. S. Huang, M. Ortega, M. Mehrotra, S. Beckman, "Relevance feedback: a power tool for interactive content-based image retrieval", IEEE Trans. on Circuits and Video Technology, 8(5), pp. 644-655, 1998.
39. S. Liang, Z. Sun, "Sketch retrieval and relevance feedback with biased SVM classification," Pattern Recognition Letters, 29, pp. 1733-1741, 2008.
40. S.D. MacArthur, C.E. Brodley, C.-R. Shyu, "Interactive Content-Based Image Retrieval Using Relevance Feedback" , 12(1), 14-26. Computer Vision and Image Understanding 88, 5575, 2002.
41. S.H. Huang, Q.J Wu, S.H. Lu, "Improved AdaBoost-based image retrieval with relevance feedback via paired feature learning.", ACM Multimedia Systems, 12(1), 14-26, 2006.
42. Y. Wu, A. Zhang, "Interactive pattern analysis for relevance feedback in multime information retrieval," ACM Journal on Multimedia Systems, 10(1), pp. 41-55, 2004.
43. D.-H. Kim and C.-W. Chung,"Qcluster: Relevance Feedback Using Adaptive Clustering for Content Based Image Retrieval," Proc. ACM Conference on Management of Data, 2003.