Multimed Tools Appl (2009) 41:125–159
DOI 10.1007/s11042-008-0222-3

Parallel neural networks for multimodal video genre classification

Maurizio Montagnuolo · Alberto Messina

Published online: 19 September 2008

Abstract Improvements in digital technology have made possible the production and distribution of huge quantities of digital multimedia data. Tools for high-level multimedia documentation are becoming indispensable to efficiently access and retrieve desired content from such data. In this context, automatic genre classification provides a simple and effective solution to describe multimedia contents in a structured and well understandable way. We propose in this article a methodology for classifying the genre of television programmes. Features are extracted from four informative sources, which include visual-perceptual information (colour, texture and motion), structural information (shot length, shot distribution, shot rhythm, shot clusters duration and saturation), cognitive information (face properties, such as number, positions and dimensions) and aural information (transcribed text, sound characteristics). These features are used for training a parallel neural network system able to distinguish between seven video genres: football, cartoons, music, weather forecast, newscast, talk show and commercials. Experiments conducted on more than 100 h of audiovisual material confirm the effectiveness of the proposed method, which reaches a classification accuracy rate of 95%.

Keywords Video annotation · Genre recognition · Neural network · Feature extraction · Multimedia semantics

Maurizio Montagnuolo is a PhD student supported by EuriX s.r.l., Turin, Italy—http://www.eurixgroup.com.

M. Montagnuolo (✉)
Department of Computer Science, University of Turin,
Corso Svizzera 185, 10149 Turin, Italy
e-mail: montagnuolo@di.unito.it

A. Messina
Centre for Research and Technological Innovation, RAI Radiotelevisione Italiana,
Corso Giambone 68, 10135 Turin, Italy
e-mail: a.messina@rai.it

1 Introduction

During the twentieth century media became very deeply ingrained in our lives, due to the rapid evolution of the information technology (IT) industry. In fact, the increasing power of electronic circuitry in personal computers, and consumer electronics, in conjunction with the decreasing cost of high-bandwidth networks, made the availability of digital content continuously increasing. Nowadays, thanks to the Internet and the digital TV, broadcasters are providing new online video services [24] and, at the same time, consumers are becoming content producers.[1]

As large-scale multimedia collections come into view, there is an urgent need for characterisation efforts on high level multimedia content descriptions, so that users can select desired contents at the semantic level, by making few simple operations. Genre annotation is the simplest way to give users the ability to select contents of interest in large collections. For example, the use of video genre recognition allows users to retrieve video contents according to classes, such as movies, newscasts, cartoons, by exploiting the idea that contents belonging to the same class share audiovisual stylistic aspects that reflect the author's intentions in producing those contents. Similarly, music genre classification would classify a piece of music based on its musical style, such as rock, pop, or orchestral music.

In this article, an architecture for automatic genre recognition of TV programmes is proposed. In particular, in order to validate this architecture, we developed an experimental framework that is able to discern between *TV commercials, newscasts, weather forecasts, talk shows, music video clips, animated cartoons* and *football match videos*. The remaining of the article is organised as follows. Section 2 reviews the state of the art in video genre classification. Our work on video classification is then introduced in Section 3. Section 4 details the set of extracted features, including both low-level visual descriptors and higher level semantic information. The process of genre classification based on parallel neural networks is described in Section 5. Experimental results are given in Section 6. Final conclusions, outlines about future work on the topic and examples of some potential end applications are proposed in Section 7.

2 State of the art in video genre classification

Editing is an important stylistic element because it affects the overall rhythm of the video document [4]. Therefore, layout related statistics are well suited for classifying video documents in pre-defined categories. This approach is known as *genre theory*.[2]

The word *genre* derives from the French word for *kind, class* or *type*. Apart from this simple definition, the meaning of genre applied to media content may be much more complex because it includes social, historical, cultural and subjective factors.

[1] Consider, for example, the YouTube video site—http://www.youtube.com/ (last accessed: May 13th, 2008), which allows users to upload, watch and share multimedia video files.

[2] http://www.aber.ac.uk/media/Documents/intgenre/intgenre.html (last accessed: November 14th, 2007)

In the television context, the meaning of *genre* is intended as the *description of what the television audience would expect to see*. Each class represents a specific video genre, e.g. movies, football matches, newscasts, cartoons, commercials, and so on. More specific genres can be associated with a general genre. For example, a programme that reviews the results of football matches would be a member of the genre *football*, which in turn would be a member of the genre *sports*.

Genre annotation is the simplest way to give users the ability to select objects of interest in large collections, although it is not a straightforward task to be performed automatically. A related problem is finding the most proper genre taxonomy, to maximise the access and selection efficiency of the system. The taxonomy may be subjective, time- and data- dependent, and must be easily understandable and browsable by the average user [29].

In this section we review the state of the art research works on video genre classification, also providing an analysis of strengths and weakness of several approaches previously proposed in the literature.

2.1 Video genre classification models

The first attempts aimed at automatic classification of video genres started in the mid of 1990s. In the 1995, Fischer et al. [16] proposed a 3-step approach based on a taxonomy of five genres: tennis, car-racing, news, commercials and cartoons. The first step involves the extraction of *syntactic properties* of the video document (i.e. the colour and audio statistics, cut detection, motion vectors and object segmentation). The second step regards the identification of *style attributes*, such as lengths and transitions of scenes, camera and object motion, use of speech, music and text. Video editors implicitly use these elements, which derive from the syntactic properties, to emphasise the video peculiarities. Thus, the last step provides the determination of the video genre, matching the *style profile* derived from the style attributes with well-known profiles, each of them typical of one particular genre.

Afterwards, several classes of statistical pattern recognition tools were used to address the genre classification problem [1, 11, 13, 17, 19, 22, 38, 42, 49, 52, 59, 61]. The approaches in [1, 17, 19, 42, 46] led to focus on only one kind of genre (either cartoons or commercials or sports).

Based on the observation that different genres show different face and text trajectories, Dimitrova et al. [11] used face and text tracking with a hidden Markov model (HMM) classifier [36] to distinguish between four TV genres. An HMM-based classifier was also used by Taskiran et al. [49] to characterise the structure of soap operas, comedies and sports video. Roach et al. [38] adopted an approach based on background camera motion, foreground object motion and aural features, to represent the dynamic content of five video genres (sports, cartoons, news, commercials and music). Classification was performed using a Gaussian mixture (GMM) [51] classifier. Experimental results denoted a correct classification rate of about 96% using video sequences of about 30 s length. Xu et al. [59] considered a set of five genres, which includes cartoons, commercials, music, news and sports. They used a principal component analysis (PCA) [21] on spatio-temporal audiovisual feature vectors and a GMM-based classifier. In [52], the same set of genres was employed in a C4.5 Decision Tree classifier [41] trained on a ten-dimensional feature vector. Liu et al. [22] used weather forecasts instead of musics and a five-states

HMM classifier. Dinh et al. [13] split the music genre into two sub-genres (music shows and concerts), thus resulting in an experimental set of six video genres. The efficiency of three different classifiers, including a decision tree, a support vector machine (SVM) [54] and a *k*-nearest neighbour (kNN) [10] classifier, was tested. Moreover, the effectiveness of the audio clip length was investigated. In [46], camera motion parameters are adopted to discern among different types of sports.

In other recent approaches, multiple classifiers [53, 57, 60] and expert systems [8, 14, 62] were also employed. Since complexity increases with the number of employed features, feature selection is a crucial step to reduce the spatio-temporal redundancy in the input data. A solution to this problem for the video genre classification task is provided in [8]. A rule-based system for automatic annotation of videos is presented in [14]. In [53], authors use a hybrid ensemble of elementary classifiers based on Markov models and support vector machines to discern among six genres, three of which are sports, and using 20 s-long clips and a total training set of 3 h. A multi-level Pseudo-2D-HMM classifier is proposed in [57] to automatically recognise the genre of sport videos. In [60], the authors use a hierarchical support vector machine architecture to distinguish among three different sets of genres taken from a genre taxonomy (generic video, sports and movies). Fuzzy classification of video sequences is proposed in [62].

As talk shows play an important role in the daily programming of the TV channels, we considered this genre in addition to those used in [13, 22, 38, 52, 59]. Finally, we would like to cite the work in [33], which describes a system based on Markov models and regression trees, to improve TV programme schedules offered by electronic program guides and TV magazines.

2.2 Detection of TV commercials

Tools for detection of TV commercials allow users to automatically recognise the presence of advertisements into broadcast programmes. This topic mainly attracts consumer electronics companies, which are aimed at developing intelligent digital video recorders that are able to detect and remove commercials from recorded programmes [9, 12, 18]. A good reference for an introductory critical analysis of the potentials and opportunities for commercials detection applications can be found in Satterwhite et al. [43].

The first attempts for commercials detection appeared between the end ’80s and the beginning ’90s. The analogical methods proposed in [3, 31] were based on the observation that some black frames (or other monochromatic frames, such as white or even blue in French broadcast channels) usually separate contiguous commercials.

In [42], commercials detection is based on the concept of *signature*. First, a set of signatures (or fingerprints) is extracted from the visual content of representative commercials frames and stored in a database. Then, a not annotated video stream is temporally segmented into shots. Finally, the visual properties of the key-frames representative of each shot are matched with the signatures in the database. The comparison of key-frames is based on colour histogram intersection. Basing on the fact that commercials are relatively rare and different compared to the main programme, Goh et al. [18] have modelled commercials detection as a binary learning problem. Two classes have been defined: the former collects *usual* events (the main programmes), while the latter collects *unusual* events (commercials). A video document is first divided into segments of fixed length. Then the K-means clustering algorithm is applied to the set of audiovisual features extracted from each segment.

In the approach proposed by Albiol et al. [1] a three-state machine is used to model the commercial detector. The first state (called *initialisation* state) occurs during the system set-up procedure. It is used to extract a binary mask that locates the position of the TV channel logo. The remaining states refer to commercials shots (*commercial* state) and to non-commercials shots (*program* state). After initialisation, the system changes to the *program* state. Hereafter, the state machine will be always in either the program or commercial state. Transitions between these two states depend on shot duration and logo presence. The first measure is related to the fact that video shots are usually shorter during commercials. The second information is useful because the logo of the TV channel is often omitted during commercials.

2.3 Recognition of animated cartoons

The word *cartoon* derives from the Italian word *cartone*, a sketch drawing executed on a paper for future drawings or satirical effects. In television and cinema productions, an animated cartoon is a movie (composed by a sequence of still drawings) made using animation techniques to produce a sensation of movement. Cartoons often present long scene duration and low camera motion energy. In addition, due to the fact that cartoons are produced in studios, and so they are not affected by environmental noise, audio tracks of animated cartoons usually show periods of no signal (zero amplitude) between noise, music or speech. Finally, colours are very often saturated and uniform. Among the few approaches addressing the topic of cartoon recognition are those in [17, 19, 37]. All these methods adopt several low-level features extracted from the aural and/or visual modality, to produce several models of cartoons and non-cartoons video sequences.

The approach published by Roach et al. [37] is based on only motion information. The experiments were carried out on a small database of approximately 20 min duration, including eight cartoons and 20 non-cartoons video sequences.

Ianeva et al. [19] investigated the use of basic visual properties of cartoons. The authors considered only individual video frames from which the following six descriptors are extracted: (1) the average colour saturation; (2) the number of pixels whose brightness is greater than 0.4; (3) the normalised 3×3×5 HSV colour histogram [45]; (4) the 8×5 normalised edge histogram, which represents the phase and the magnitude components of the gradient of the luminance channel; (5) the compression ratio, which indicates the frame complexity; and (6) the texture granularity, which denotes the dimensions of the objects represented in the frame. This feature vector is used as input of a support vector machine. The experiments were conducted using a set of 24,000 JPEG images from the TREC-2002 video track test collection,[3] achieving an error rate of about 25%.

Glasberg et al. [17] adopted an approach based on both audio and video modalities. The audio modality include 13 static Mel frequency cepstral coefficient features

[3] Available online at: http://www-nlpir.nist.gov/projects/t2002v/t2002v.html (last accessed: September 14th, 2007).

and their first and second derivatives, resulting in a 39-dimension feature vector. The video modality include five descriptors representing colour properties (saturation and nuance), grey scale levels (brightness), edge characteristics, and motion activity, resulting in an eight-dimensional feature vector. Audio and video features are modelled by a mixture of Gaussians and a multilayer perceptron, respectively. Classification is then performed considering a combination of the outputs of these two classifiers. The experiments were performed on a video database of 200 min length, achieving a total correct classification rate of about 90%.

2.4 Discernment of sports, music, news, cartoons, and commercials

A pioneer work whose goal is to discern between more than a single genre is that of Fischer et al. [16], who manually analysed the rules used by video editors (e.g. shot lengths, or volume of audio tracks) to derive stylistic patterns. Several classes of pattern recognition tools and artificial intelligence techniques have been used to address the problem so far.

In [52], the authors define a set of nine visual features that reflects the human perception in the discernment of video genres, resulting in a ten-dimension feature vector. Features f_1 and f_2 (average shot length and percentage of abrupt and gradual shot transitions, respectively) are related to the editing effects used. Features f_3 to f_6 (camera motion energy, pixel luminance variance, quite scenes rate, average motion runs length) are related to the camera motion properties of the video sequence. Finally, features f_7 to f_9 (average standard deviation of luminance, percentage of high brightness pixels and percentage of high saturation pixels) are related to the colour content of the video. A genre classifier based on a C4.5 decision tree [35] trained on these features is employed. Experiments are conducted on about 8 h of TV programmes, coded in MPEG1. The authors studied the goodness of the proposed features set as well as the optimal length of a clip needed to recognise its genre. Video clips of 60 s length get the best experimental results (correct classification rate of 86.2%, 81% and 83.1% for the best, worst and average case, respectively).

Xu and Li [59] have investigated the achievement of automated video genre classification using compact representations of low-level aural-visual features. The authors aimed at overtaking the *"curse of dimensionality"* problem,[4] using principal component analysis on spatial-temporal audiovisual feature vectors. The first 14 Mel Frequency Cepstral Coefficients are used as spectral features to represent the audio signal of a video clip. The MPEG-7 colour and texture descriptors are used as visual content representative features. Finally, energy motion is described by the mean and standard deviation of the magnitudes of MPEG video motion vectors. Classification is performed by adopting a Gaussian mixture classifier. Those features are extracted from short videos of duration W between 2 and 20 s. According to the authors, the goodness of the proposed approach depends on many factors, including the number of principal components (P) and the number of Gaussian components (K). The authors' experiments denote that the classification accuracy augments with the increase of W and P, and shows the best results for K between 3 and 10.

[4] Richard Bellman coined the term *"curse of dimensionality"* to describe the difficulty of evaluating Probability Density Functions on high-dimensional feature spaces [2].

Liu et al. [22] employ a five-states hidden Markov model to discriminate five types of TV programmes, namely commercials, basketball games, football games, news reports and weather forecasts. Audio features are used to characterise each genre. Experimental results were obtained using 20 min video for each genre from several TV channels. The system achieves a classification accuracy of about 93%, getting an improvement of about 12% with respect to the use of the same dataset and a neural network classifier [23].

Roach et al. [39] adopt an approach based on background camera motion and foreground object motion, to represent the dynamic content of video. Background camera motion is expressed in terms of camera pans, tilts and zooms. Foreground object motion is estimated by subtracting background camera motion from global motion vectors. These video dynamics are extracted, processed and applied to classify three genres: sports, cartoons and news. Classification is performed using a GMM-based classifier. Experimental results denote a correct classification rate of about 96% using video sequences of about 30 s length. An extended version of this work includes consideration of more genres and additional audiovisual features [38].

Dinh et al. [13] propose the use of Daubechies wavelet transform with six levels of decomposition to analyse audio tracks of video documents. The efficiency of three different classifiers, including a decision tree, a support vector machine and a *k* nearest neighbour, was tested. Moreover, the effectiveness of the audio clip length was investigated. Experiments denote that the kNN-based classifier obtains the best classification results. In addition, there are not significant differences in the classification accuracy when using audio clips of different lengths.

2.5 Discernment of news, movies, sports and commercials

Movie genre includes a great variety of sub-genres, such as comedy, soap, sitcom and action. The method proposed in [11] is based on the observation that different TV categories show different text and face trajectory patterns. Face and text tracking is applied to video segments to extract trajectories of faces and text. Then, four hidden Markov models are implemented and trained to recognise four video classes: news, commercials, sitcom and soap. The authors report an overall classification accuracy of about 85% on 1 min length video clips. An HMM-based classifier is also used by Taskiran et al. [49] to characterise the structure of soap operas, comedies and sports video. In [61], a decision tree classifier is used to classify TV programmes as different genres, including music video, commercials, movies (love, comedy, ethical, tragedy, action) and sports. The same classifier and the same set of genres plus news is also used in [60].

2.6 Discussion

Semantic multimedia classification is a key issue in modern multimedia applications, where the number of available items is dramatically growing and there is an increasing demand for multimedia objects fruition in distributed scenarios. In this context, selection by genre is a simple and effective mechanism for most of the users interested in these applications. The fundamental assumption in this domain is that *multimedia producers very often work according to some common stylistic patterns*. They use specific and recurring aggregations of low-level perceivable features (e.g.

colours, editing effects), of structural properties (e.g. rhythm in the sequence of shots) and of cognitive properties (e.g. shots representing softly smiling people, or natural landscapes) to communicate some high-level concepts (e.g. peace and relaxation). Concepts are further combined, at a higher level, to produce more general concepts, which can result in the genre of a video (e.g. sports, news).

From the analysis of previous work, some considerations can be made. First of all, despite the number of existing efforts in the field, the problem of comprehensively discerning broadcast programmes by genre has not been satisfactorily solved yet. This depends on two concomitant factors: the difficulty to analytically define the *concept of genre*, which is a typical subjective and domain dependent concept, and the intrinsic multiformity of TV programmes, which can simultaneously belong to several genres. Because of this, most of the existing approaches either attempted to discern only few genres from a simple taxonomy, or to distinguish a single well-defined genre from all the others, or only focussed on one very specific domain, such as sports video.

Another common drawback of existing works is the relatively modest dimension of the learning database (usually from few minutes to some hours, with the exception of [60], where the authors use a database of more than 60 h). By using small datasets the risk is to produce highly data-coupled classification models and poor generalisation capabilities in real scenarios [58]. Furthermore, in literature the learning sets are not well-contextualised, i.e. very often they are declared to be random collections of objects selected from a variety of delivery channels. In our opinion this may introduce unneeded complexity in classification models, which try to tackle the huge variety of editorial languages that can be encountered over time and through media channels, and that can all be generically assigned to the same genre. To overcome these limitations we developed a learning database counting more than 100 h of programmes taken in a controlled period of time and from an affine set of delivery channels. Genre annotations were taken from official schedule annotations, i.e. we did not annotate the items ourselves. In addition, we used complete broadcast programmes (e.g. an entire talk show), thus limiting potential bias caused by authors in the clip selection phase.

A third problem deals with the kind of classifier employed. In many cases, classical statistical pattern recognition methods (i.e. crisp clustering algorithms [13, 42], decision trees [13, 52], support vector machines [13, 19, 60], Gaussian mixture models [38, 59] and hidden Markov models [1, 11, 22]) were used. The biggest problem behind statistical pattern recognition classifiers is that their efficiency depends on the class-separability in the feature space used to represent the multimedia data. We addressed this problem by using four parallel neural networks, each specialised in a particular aspect of multimedia contents, to produce more accurate classifications. In addition, neural networks do not require any a priori assumptions on the statistical distribution of the recognised genres, are robust to noisy data and provide fast evaluation of unknown data.

Another important factor in the development of a classification system is the choice of the data validation strategy. Most of the earlier works used hold-out validation (HOV) of data [11, 13, 16, 17, 22, 42, 52, 59–61], whereby the experimental dataset is randomly split into two sub-sets for training and testing. The main drawback of this method is that the classification accuracy may significantly vary depending on which sub-sets are used. One approach used leave-one-out cross validation (LOOCV) of data [38], which involves the recursive use of a single item from the experimental dataset for testing, and the remaining items for training. This method is time consuming and its classification accuracy may be highly variable. In our system, we chose to use K-fold cross validation (KFCV) of data. This approach limits potential bias that could be introduced in the choice of training and testing data.

Final considerations regard the work by Poli and Carrive [33]. This approach is quite different from our work for at least two reasons. First, the nature of the considered data is considerably different. In [33], the authors consider live multimedia streams, while in our work the data set is composed by a collection of distinct archived multimedia objects. Second, the approach in [33] is channel- and time-dependent, as it needs the knowledge about previous TV programme schedules in order to build the Markov model. In contrast, our approach has been demonstrated to be channel- and time-independent (see Section 6), and it does not need knowledge about how programmes have been arranged in the programming schedule. Furthermore, our approach is able to classify also programmes that have not been published yet, and this constitutes an interesting feature for strategic programme planning for a broadcaster.

3 Overview of the TV genre modelling and classification process

This section briefly introduces the functionality of the blocks composing our architecture. Further detail on all of them will be provided in Sections 4 and 5.

The proposed system makes use of artificial intelligence (i.e. parallel neural networks) to approximate human-like reasoning and to derive semantic information from multimedia (i.e. broadcast video genre). In designing our system we have borrowed experience from previous research focussing our attention on the scalability and efficiency of the system as well. The proposed approach results in a modular structure in which features are extracted, represented and matched independently, according to different extraction methods and representation metrics.

The process of TV genre modelling and classification is shown in Fig. 1. In order to train the system and validate the prediction model, each TV programme was first manually pre-labelled according to the following schema, which is inspired by the scheme commonly used to classify radio and TV programmes:

$$\text{C:S:D:Tp:M}$$

where:

- C indicates the *context*, i.e. the data that is referred to in the label, such as “content data”, “scheduling data”, “transmission data”, “financial data”;
- S denotes the reference metadata *schema*, e.g. the ESCORT (EBU System of Classification Of Radio and TV Programmes) [15], the TV-Anytime metadata schema [32], the NLog library;[5]

[5] Available online at: http://www.nlog-project.org (Last accessed: September 28th, 2007).

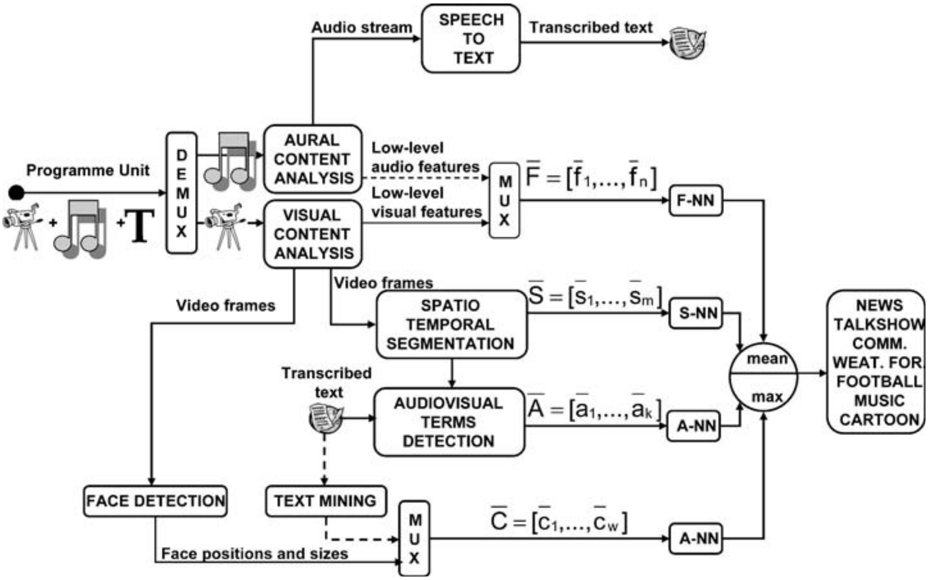


Fig. 1 The proposed architecture for automatic TV genre classification. Given an input video sequence, we separate the video and audio tracks from each other. Visual content analysis extracts information concerning perceptual ($\bar{F}$), structural ($\bar{S}$) and cognitive ($\bar{C}$) properties of the video. Aural content analysis extracts information concerning acoustic ($\bar{A}$) properties of the video. Additionally, a speech to text engine processes the audio track to produce a transcription of the spoken words (T) that is used to enrich the set of acoustic features. These feature vectors serve as the input for a parallel neural network system that performs classification of seven video genres. Dashed lines indicate future development, which might include e.g. more acoustic features, such as temporal features and spectral analysis, and text mining techniques

- D is related to the *dimension*, i.e. descriptive information of the programme, such as “genre”, “intention”, “intended audience”;
- Tp indicates the *type* of the programme (w.r.t. the kind of dimension), such as “news”, “science”, “football” in case of the “genre” dimension;
- M denotes the *degree of membership* of the programme to the specified type.

An example of label adopted for the newscast programmes is the following:

_:_:genre:news:1.0

where the character ‘_’ indicates the default value. This label is utilised to identify the programmes used to train the genre classifier, and to validate the classification results.

Then, we associate to each programme a *hierarchically structured surrogate*, i.e. a feature vector that is a combination of several sub-parts, each representing a particular aspect of multimedia content. The surrogate contains data extracted by processing the audiovisual content, ranging from low-level perceptual descriptors, such as colours, textures and motion, to higher-level, human-centred features such as faces, structures and transcribed texts. Many details about the extracted features will be given in Section 4.

The surrogate is treated as input for a neural network. The main neural network is composed by four parallel sub-networks. Each sub-network is specialised in a particular part of the surrogate, and it works independently from the others. The outputs of all individual sub-networks are finally combined by an ensemble method to get the final classification. The ensemble neural network is thus based on the classification results derived from the analysis of each aspect of the examined video genres.

3.1 The aural and visual content analysis modules

The construction of a genre classification model requires the extraction of a numerical representation of multimedia content. We call this numerical representation as the *programme surrogate* (PS). The programme surrogate is a *multimodal* vector $\mathbf{PS} = (\mathbf{F}, \mathbf{S}, \mathbf{C}, \mathbf{A})$ that integrates four monomodal information sources, each representing a particular aspect of multimedia content [27]: **F** represents the features at the visual low-level. **S** carries the structural features; **C** carries cognitive features; finally, **A** contains the aural features.

3.2 The spatio-temporal segmentation module

The spatio-temporal segmentation module performs visual shot detection and clustering. We define *shot* a sequence of contiguous frames characterised by similar visual properties. Analogously, a *shot cluster* is a group of shots sharing similar visual content. The spatio-temporal segmentation module is based on an optimised bottom-up clustering method that adopts histogram intersection as the distance metric between feature histograms. The segmentation process is performed in 5 steps:

1. *Feature histograms extraction*. Video frames are firstly divided into N_b 16×16 pixel blocks. Then, seven low-level visual descriptors (i.e. hue, saturation, value, luminance, contrast, directionality and temporal activity) are extracted from each block [28]. All features are represented using a 65-bin histogram, where the last bin is used to count the fraction of pixels whose measurement returns a undefined value (e.g. hue for grey pixels). At this step, each frame is thus represented by $7 \cdot N_b$ *local* histograms, one for each block and each feature;
2. *Preliminary shot detection*. In order to initialise the shot clustering algorithm, a coarse shot detection is performed as second step of the segmentation process. The core algorithm consists of an adaptive threshold shot detection evaluation. For each frame, its temporal activity *global* histogram is built, summing up each of the N_b local histograms. Two consecutive frames are considered to be members of the same shot if the following condition is verified:

   $$d^{(tmpact)}(h_i, h_{i+1}) \leq \alpha \tag{1}$$

   where h_i is the temporal activity histogram of the frame i and α is a fixed threshold.
3. *Partial clustering on fixed-length segments*. The whole video is firstly split into K uniform segments of length $l = N_f/K$, where N_f is the total number of frames in the video. The clustering algorithm is then performed on each segment, using the output of the preliminary shot detection as the initialisation condition, and the histogram intersection as the core distance. After this step, each frame is

labelled with the tuple (k, γ), $k = 1, \ldots, K$ and $\gamma = 1, \ldots, \Gamma_k$, where k and γ are, respectively, the segment number and the cluster number to which the frame belongs, and Γ_k is the total number of clusters found in the kth segment;

4. *Selection of relevant clusters found in the partial clustering phase*. All irrelevant clusters found in the previous step are filtered out. The filtering method adopted is based on the concept of *absolute cluster membership*: all clusters that include single shots with duration less than a fixed parameter threshold are rejected;
5. *Re-clustering on selected clusters*. The selection step produces a reduced set Γ' of clusters. A final clustering, restricted to the frames belonging to the clusters included in Γ', is then performed. After this step, each shot s_i is finally associated to a tuple $\ell^{s_i} = (type, start, end)$. The *type* parameter can take values in $\{stable, transitional\}$ depending on the type of editing effect applied to separate two consecutive shots (i.e. direct cuts or other abrupt transitions denote *stable* shots, while fades, dissolves and other non-abrupt transitions denote *transitional* shots). The *start* and *end* parameters identify the starting and ending point of s_i in the programme timeline.

3.3 The face detection module

The face detection module implements face detection techniques.[6] Automatic face detection aims to determine locations and sizes of human faces in digital images. Earlier face detection techniques focused on the detection of only frontal human faces. A survey of these methods can be found in [7]. More recent algorithms attempt to solve the more general and difficult problem of multi-view faces (i.e. rotated faces) detection.

3.4 The audiovisual terms detection module

The audiovisual terms detection module performs a statistical analysis to find out the frequency of some peculiar characteristics of audiovisual works carrying semantic information (e.g. the average number of words per second, the percentage of music in the audio track, the number of speakers).

3.5 The parallel neural network

The parallel neural network module is used as a first proof-of-concept application to assist a human television archivist in the classification and annotation process. It uses an ensemble neural network as core classifier. The global neural network is composed by four parallel sub-networks, i.e. four 3-layer perceptrons with Sigmoid activation functions.

Neural networks are powerful in realising complex non-linear problems. In addition, they do not require any a priori assumptions on the statistical distribution of the recognised genres, they are robust to noisy data and they provide fast evaluation of unknown data.

[6] See http://www.facedetection.com to find many available resources about the face detection task (last accessed: October 10th, 2007).

In our approach, each sub-network is specialised in a particular information source, and it works independently from the others. The ensemble neural network is thus based on the classification results derived from the analysis of each aspect of the examined video genres. In our method, it is easy to consider either a new feature or a new information source, by simply including an additional set of input neurons to the correspondent sub-network, or by adding a new sub-network to the global ensemble prediction system. Thus, it is needed to retrain only the sub-network with the new feature, without retraining the other sub-networks.

4 Proposed feature sets

Effective multimedia indexing and retrieval requires a multimodal approach, considering aural, visual, and textual information altogether [44]. Our system is multimodal in that it uses a surrogate vector to represent the set of features extracted from all available media channels included in a multimedia object. These media channels are representative of modality information (both visual and aural), structural-syntactic information and cognitive information. A introductory analysis concerning the representation of multimedia information content of audiovisual material can be found in [29].

Starting from the basic media types introduced above, we derive the TV programme surrogate vector $\mathbf{PS} = (\mathbf{F}, \mathbf{S}, \mathbf{C}, \mathbf{A})$. This vector collects four sets of features that capture visual (**F**), structural (**S**), cognitive (**C**) and aural (**A**) properties of the video content, as reported in Table 1. We have originally designed some of these features to reflect the criteria used by editors in the multimedia production process. Some of these features are well defined in literature (e.g. the average shot length and the average speech rate), other ones are new (e.g. the shot cluster duration

Table 1 Overview of the programme surrogate descriptors

PS part	Name	Notation	Dimensionality
F	Hue	$\mathbf{H}$	30
	Saturation	$\mathbf{S_{at}}$	30
	Value	$\mathbf{V}$	30
	Luminance	$\mathbf{Y}$	30
	Contrast	$\mathbf{C_{on}}$	30
	Directionality	$\mathbf{D_{ir}}$	30
	Temporal activity	$\mathbf{T_{em}}$	30
S	Shot length distribution	$\mathbf{SLD}$	9
	Shot temporal activity	$\mathbf{STA}$	5
	Average shot length	ASL	1
	Shot cluster duration	CLD	1
	Shot cluster saturation	CLS	1
C	Face number distribution	$\mathbf{FAD}$	11
	Face position distribution	$\mathbf{FPD}$	9
	Face covering percentage	FCP	1
	Average face number	AFN	1
A	Audio segmentation analysis	$\mathbf{ASA}$	4
	Background audio analysis	$\mathbf{BAA}$	3
	Average speech rate	ASR	1

and saturation, the face position distribution and the shot temporal activity). The four feature sets included in the programme surrogate are detailed in the following subsections.

4.1 The low-level visual part of the programme surrogate

The low-level visual component of the programme surrogate includes $N_F = 7$ features. Colours are represented in the **HSV** (Hue, Saturation, Value) colour space because of its perceptual uniformity and its similarity to the human vision system [45]. Luminance ($\mathbf{Y}$) is represented in a grey scale in the range [16, 233], with black and white corresponding to the minimum and maximum value, respectively. Tamura's features [47] are used to describe contrast ($\mathbf{C_{on}}$) and directionality ($\mathbf{D_{ir}}$) of textures. Temporal activity information ($\mathbf{T_{em}}$) is based on the displaced frame difference (DFD) [50] for window size $t = 1$.

First, as the low-level audiovisual features are natively computed on a frame by frame basis (i.e. there is one hue histogram for each frame of the programme), a global histogram for each shot detected by the segmentation module is provided. Then, we model each of these histograms by a K-component Gaussian mixture with $k = 10$ [28]. The ith, $i = 1, \ldots, 10$ component of the mixture is a couple $< \omega_i, N(\mu_i, \sigma_i) >$, where w_i is the weight of the component and $N(\mu_i, \sigma_i)$ is a Normal distribution with mean μ_i and variance σ_i^2. Each feature of each shot is thus represented by the following $3K$-dimension vector:

$$\theta_i^{(f)} = \{\omega_1, \ldots, \omega_K, \mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2\}. \tag{2}$$

We averaged all of the $\theta_i^{(f)}$ vectors on the total number of shots to obtain a single characteristic vector $\theta_m^{(f)}$ for each feature. Low-level visual features are finally represented by the vector $\Theta = (\theta_m^{(1)}, \ldots, \theta_m^{(N_F)})$, whose size depends only on the number of Gaussians K. The optimal number of K was experimentally determined to be equal to 10. Figure 2 shows an example in which a 65-bin hue histogram is fitted by a ten-component Gaussian mixture.

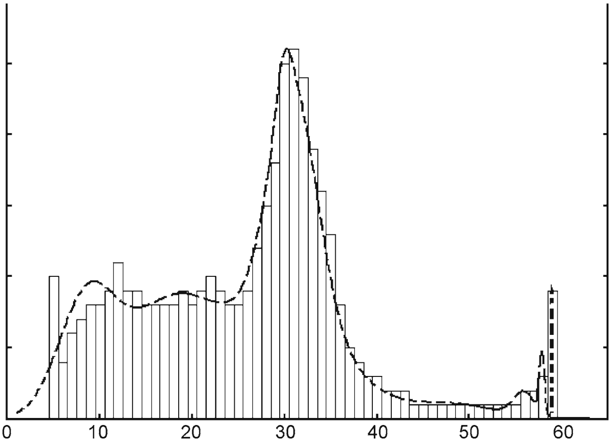


Fig. 2 Example of a hue histogram fitted by a ten-component Gaussian mixture

4.2 The structural part of the programme surrogate

The second information source $\mathbf{S} = (ASL, CLD, CLS, \mathbf{SLD}, \mathbf{STA})$ includes five descriptors of the video's *structural* properties. ASL is the average shot length; CLD and CLS are, respectively, the shot cluster duration and saturation; **SLD** is the shot length distribution; **STA** is the shot temporal activity.

4.2.1 Average shot length

The average shot length indicates the average duration of the shots (measured in seconds) included in the programme. This feature was first proposed by Vasconcelos et al. [55]. The average shot length captures information about the average rhythm of the video. It is calculated by averaging the duration of the shots detected by the shot detection algorithm:

$$ASL = \frac{1}{F_r N_s} \sum_{i=1}^{N_s} \Delta l_i, \tag{3}$$

where F_r is the frame rate of the video (i.e. 25 f.p.s.), N_s is the total number of shots in the programme and Δl_i is the shot length, measured as the number of frames within the ith shot. Our experimental observations showed that cartoons and commercials are characterised by short duration shots. In contrast, weather forecasts and talk shows typically contain longer shots.

4.2.2 Shot cluster duration and saturation

The shot cluster duration (CLD) is the normalised duration (w.r.t. the total duration of the programme) of visual shots taking part in a shot cluster containing at least two elements. Analogously, the shot cluster saturation (CLS) is the ratio between the number of visual shots aggregated in shot clusters counting at least two elements and the total number of shots included in the programme:

$$CLD = \frac{\sum_{c=1}^{C} \sum_{i=1}^{N_c} \Delta d_i^{(c)}}{N_f}, \forall c : N_c \geq 2 \tag{4}$$

$$CLS = \frac{\sum_{c=1}^{C} N_c}{N_s}, \forall c : N_c \geq 2 \tag{5}$$

where: C is the total number of detected clusters in the programme; N_c is the number of shots included in the cth cluster; $\Delta d_i^{(c)}$ is the duration (measured in frames) of the ith shot in the cth cluster; N_f is the total number of frames within the programme; and N_s is the total number of shots within the programme. Both CLD and CLS get values in the range 0 to 1.

As an example, Table 2 reports the average values of the CLD and CLS obtained for the seven considered television genres. It shows, for example, that both the cluster duration and the cluster saturation for weather forecast programmes approach to zero, due to the fact that there are few repetitive shots in these programmes. On the other hand, as talk shows have many repetitive shots, which include e.g. interviews or monologues, they are usually characterised by higher cluster saturation and duration values.

Table 2 Average values of shot cluster duration and saturation for the seven considered TV genres (Unit: 1)

	Comm.	News	Weat.For.	Cart.	Music	T.Show	Foot.
CLD	0.142	0.299	0.081	0.212	0.251	0.629	0.438
CLS	0.147	0.146	0.089	0.143	0.229	0.531	0.258

4.2.3 Shot length distribution

The shot length distribution describes how shots lengths are distributed along the video. As different video genres have different editing rules and patterns, the distributions of shot lengths can be used as discriminant features to analyse the video content.

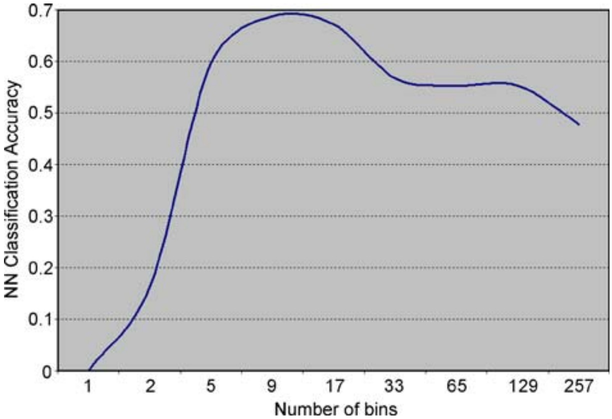
Various approaches have been proposed to model the distribution of shot lengths, using either probability distributions or histogram-based models [6, 48]. In this work, the shot length distribution is modelled by a nine-bin histogram normalised by the number of total shots. Bins 1 to 8 are uniformly distributed in the range [0,30 s]. The ninth bin counts the fraction of shots whose duration is greater than 30 s.

The optimal number of bins to represent the shot length distributions was determined empirically by measuring the classification accuracy of a multilayer perceptron with one hidden layer and 16 hidden neurons. The main problem in selecting the optimal number of bins consists in balancing the increased classification accuracy with the increased storage and computation costs, resulting from the insertion of additional bins. Figure 3 shows that the best classification accuracy was achieved for $n \approx 9$.

4.2.4 Shot temporal activity

The shot temporal activity describes how the shots are distributed along the programme timeline. **STA** is a cumulative function $F_{sta} : [0, 1] \rightarrow [0, 1]$. Given a programme u and a fraction of its duration $t = \frac{f_i}{N_f}, \quad i = 1, \ldots, N_f$, $F_{sta}(t; u)$ is the fraction of shots occurred before t. Analytically, the temporal activity distribution is modelled by a histogram with five uniformly distributed bins. As for the shot length

Fig. 3 Classification accuracy versus the number of bins in the **SLD** histograms



distribution, the optimal number of bins used to represent the distribution of the shot temporal activity was determined empirically.

4.3 The cognitive part of the programme surrogate

The cognitive part of the programme surrogate models face-related features. At first, face detection is applied to locate the position and size of faces within the video.[7] Then we extract three features obtaining the 22-dimension vector $\mathbf{C} = (\mathbf{FPD}, FCP, \mathbf{FAD})$.

4.3.1 Average face number

The average face number measures the face-per-frame rate. It stems from the observation that different TV genres show different amounts of faces. Analytically, the average frame number is computed as the ratio between the total number of faces detected in the programme (N_{faces}) and the programme duration (D) measured in frames:

$$AFN = \frac{N_{\text{faces}}}{D} \tag{6}$$

Our experimental observations showed that the average face number for weather forecast programmes approaches to zero, due to the fact that the majority of these programmes contain only graphics and text. On the other hand, as in talk shows there are many people who are speaking or listening (i.e. the presenter, the guests and the audience), these programmes are usually characterised by higher (w.r.t. the other genres) AFN values.

4.3.2 Face number distribution

The face number distribution (**FAD**) describes the distribution in the programme of the number of faces appearing in the same frame. This distribution is expressed by a normalised (w.r.t. the number of frames) eleven-bin histogram. The ith, $(i = 0, \ldots, 9)$ bin counts the fraction of frames containing i faces. The 11th bin counts the fraction of frames depicting at least ten faces. Even if the majority of TV programmes show no more than two or three faces, we believe that this feature is meaningful to distinguish the cases where many faces appear many times (e.g. football matches or talk shows with audience participation).

4.3.3 Face covering percentage

The face covering percentage is the covering percentage of faces in the programme, calculated as the ratio between images area containing faces and the total images area of the programme. The face covering percentage stems from the observation that

[7] Using the Fraunhofer IIS Real Time Face Detector tool, http://www.iis.fraunhofer.de (last accessed: March 28th, 2008).

Table 3 Mean values of the face covering percentage for the seven considered television genres (Unit: 100%)

Comm.	News	Weat.For.	Cart.	Music	T.Shows.	Foot.
0.66	1.729	0.245	0.156	0.884	3.896	0.071

different TV genres show different face dimensions. Analytically, the face covering percentage is defined by the following equation:

$$FCP = \frac{100}{D \cdot W \cdot H} \sum_{i=1}^{N_f} (w_i \cdot h_i) \tag{7}$$

where: N_f is the total number of faces detected in the programme; w_i and h_i are, respectively, the width and height (measured in pixels) of the ith face region; D is the programme duration, measured in frames; W and H are the frames dimensions.

As an example, Table 3 shows that the *football* genre presents the lowest values of the FCP (w.r.t. the other genres). This is intuitively true because football matches typically contain long field camera shots or crowded scenes (i.e. smaller face regions). On the other hand, talk show and news programmes typically contain one or more persons in the foreground of the picture. Consequently, the depicted faces are bigger (w.r.t. the frame size) and thus the face covering percentage increases.

4.3.4 Face position distribution

The face position distribution describes how faces are positioned in the programme timeline. Analytically, it is represented by a normalised (w.r.t. the total number of detected faces) nine-bin histogram. The ith bin counts the fraction of faces in the ith position in the frame. Given the Cartesian plane with origin in the central pixel of the frame, the possible positions are the following:

- *Top-right* for faces entirely included in the first quadrant;
- *Top-left* for faces entirely included in the second quadrant;
- *Bottom-left* for faces entirely included in the third quadrant;
- *Bottom-right* for faces entirely included in the fourth quadrant;
- *Left* for faces positioned across the second and third quadrant;
- *Right* for faces positioned across the first and fourth quadrant;
- *Top* for faces positioned across the first and second quadrant;
- *Bottom* for faces positioned across the third and fourth quadrant;
- *Centre*, otherwise.

4.4 The aural part of the programme surrogate

The last information source is extracted by the analysis of the programme's audio track. First, a speech to text engine processes the audio stream, using a tool based on [5]. Then, we use this information to extract three acoustic features, resulting in the feature vector $\mathbf{A} = (\mathbf{ASA}, \mathbf{BAA}, ASR)$.

4.4.1 Average speech rate

The timing of speech within audio material associated with video data is known to be variable, depending on many factors, including lexical and syntactic features, as well as emotional and cognitive aspects [56].

Similarly to the average shot length, we employ the average speech rate to capture information about the rhythm of the *spoken* parts in the programme's audio track. Analytically, this information is calculated by averaging the number of transcribed words over the duration of the programme (measured in frames).

Experimental observations showed that the words-per-second rate for football matches and cartoons are approximately the double of those for weather forecasts, talk shows and newscasts. As there are only a few spoken parts in the music contents, the average speech rate for the music class is close to zero.

4.4.2 Audio segmentation analysis

Audio segmentation analysis (**ASA**) deals with the problem of segmenting a continuous audio stream in terms of acoustically homogeneous segments. Here, this information is represented by a normalised (w.r.t. the number of audio samples) four-bin histogram, counting the percentage of *speech, silence, noise* and *music* within the programme's audio track.

As an example, our experiments showed that, as expected, in newscasts and talk shows most of the audio samples belong to the speech class (i.e. the first bin of the **ASA** histogram), while the football genre is characterised by much noise (i.e. the last bin of the **ASA** histogram), due to the shouting crowd in the audio background.

4.4.3 Background audio analysis

BAA (background audio analysis) provides information about the percentage of *silence, noise* and *music* within the spoken parts in the programme. **BAA** is modelled by a normalised (w.r.t. the number of audio samples labelled as speech content) three-bin histogram.

Empirical observations showed that, according to our intuitive expectations, the audio tracks of commercials and weather forecasts mainly contain speech plus music contents, while the dominant class for the football genre is speech plus noise (i.e. the commentator voice plus the voices of the spectators in the stadium).

5 The parallel neural network classifier

Artificial neural networks are a classification technique inspired by the human brain. One of the most popular neural networks used to solve classification problems is the multilayer perceptron (MLP), originally proposed by Rumelhart et al. [40]. This kind of neural network consists in a *input* layer, followed by one or more *hidden* layers and an *output* layer. Due to its approximation and generalisation capability, the multilayer perceptron is well suited to model any continuous function $f : \Re^N \rightarrow \Re^C$, where N is the dimension of the input space and C is the number of discriminating classes.

In this work, four parallel sub-networks (i.e. four independent multilayer perceptrons) are used to model each aspect of multimedia content, as already shown in Fig. 1. The outputs of these networks are combined together to classify the programme according to the seven considered TV genres. Thus, the genre classification problem is brought back to a multi-class supervised classification problem.

Let p be a TV programme to be classified and $\Omega = \{\omega_1, \ldots, \omega_{N_\omega}\}$ be the set of available genres. Each sub-network outputs the vector:

$$\Phi^{(p,n)} = (\phi_1^{(p,n)}, \ldots, \phi_{N_\omega}^{(p,n)}), \quad n = 1, \ldots, 4. \tag{8}$$

where $\phi_i^{(p,n)}$ is the output of the ith neuron in the output layer of the network n. Intuitively, $\phi_i^{(p,n)}$ is the membership value of p to the genre i, according to the information source n.

The outputs of the 4 sub-networks are finally combined in the vector $\Phi^{(p)} = \{\phi_1^{(p)}, \ldots, \phi_{N_\omega}^{(p)}\}$, whose elements are calculated applying the mean-rule:

$$\phi_i^{(p)} = \frac{1}{4}\sum_{n=1}^{4} \phi_i^{(p,n)}. \tag{9}$$

Many studies have been conducted on the choice of the best ensemble generation method. We chose to use the mean-rule because of its simplicity and consistent performance over a broad spectrum of applications [34]. The final classification $\omega^{(p)}$ is given selecting the genre i corresponding to the maximum element of $\Phi^{(p)}$:

$$\omega^{(p)} = \arg\max_i(\phi_i^{(p)}). \tag{10}$$

5.1 Determination of the network architecture

The development of a MLP-based model requires the definition of the network architecture. That is, the total number of layers, the number of neurons in each layer and the number of connections between neurons have to be fixed. If the network is designed to be more complex than the optimum size (i.e. there are too many layers, neurons and/or connections in the network), it will overfit the training data, thus resulting in poor classification accuracy for those patterns on which the network was not trained. On the other hand, if the network becomes too small (i.e. there are too few neurons and/or connections in the network), it will not be able to model adequately all the properties of the training data and, again, it will result in low classification accuracy.

Here, the optimal network architectures were determined through an iterative evaluation process. Each sub-network was trained on a random subset of the experimental dataset by the iRprop algorithm [20] with increasing complexity, starting from the simplest network topology and stopping the process if either the desired error ε was achieved or the maximum number of epochs M_s was reached. Then, in order to define the most suitable network architecture for each part of the programme surrogate (i.e. the number of layers, neurons and connections), we considered the following two factors:

1. The *training efficiency* η, inspired by the F-measure used in Information Retrieval and expressed as:

$$\eta = \frac{2\alpha_t(1-\beta_t)}{\alpha_t + (1-\beta_t)} \tag{11}$$

Fig. 4 Training efficiency for the aural MLP

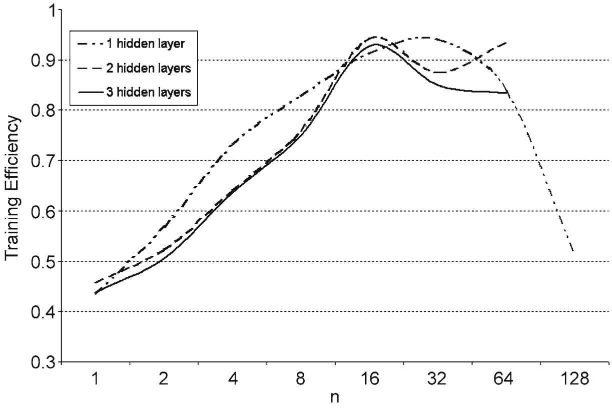


Fig. 5 Training efficiency for the structural MLP

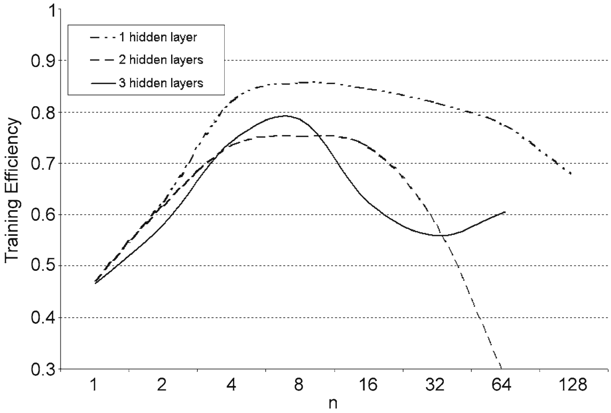


Fig. 6 Training efficiency for the cognitive MLP

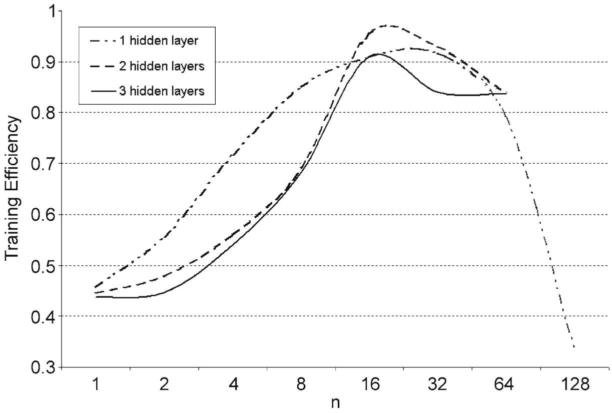
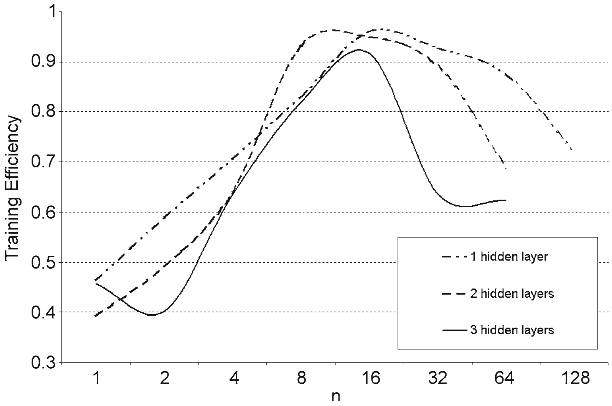


Fig. 7 Training efficiency for the visual-perceptual MLP



The training efficiency combines the training accuracy α_t (i.e. the ratio of correct items to the total number of items in the training set) and the training quality β_t (i.e. the square error between the desired output of an output neuron and the actual output of the neuron, averaged by the total number of output neurons). The training efficiency, the accuracy and the quality are all included in the range [0,1]. Figures 4, 5, 6 and 7 show the training efficiency for each part of the programme surrogate;

2. The total number of hidden neurons (HNs) and the total number of hidden layers (HLs).

Table 4 Training accuracy (α_t), training quality (β_t) and training efficiency η computed for different concatenations of the selected features and for different number of total hidden layers and total hidden neurons (Unit: 1)

Surrogate features	HLs	HNs	α_t	β_t	η
All visual features	3	16	0.996	0.015	0.991
$\mathbf{H}, \mathbf{S_{at}}, \mathbf{V}$	3	16	0.989	0.027	0.981
$\mathbf{Y}, \mathbf{C_{on}}, \mathbf{D_{ir}}$	3	16	0.834	0.093	0.868
$\mathbf{V}$	3	16	0.762	0.120	0.817
$\mathbf{C_{on}}$	3	16	0.727	0.143	0.786
All structural features	3	64	0.949	0.0935	0.927
$ASL, CLS,$ **SLD**	3	64	0.804	0.330	0.730
$ASL, CLD,$ **STA**	3	64	0.793	0.330	0.726
$AFN,$ **FAD**, **FPD**	3	32	0.991	0.017	0.987
FPD, **FAD**	3	32	0.991	0.020	0.985
All cognitive features	4	32	0.955	0.023	0.965
$AFN,$ **FAD**, **FPD**	3	16	0.879	0.060	0.908
FPD, $FCP,$ **FAD**	3	16	0.884	0.202	0.839
All aural features	3	32	0.962	0.070	0.945
ASA, **BAA**	3	32	0.952	0.083	0.933
All aural features	5	48	0.900	0.042	0.927
ASA, ASR	3	16	0.806	0.078	0.859
ASA	3	16	0.790	0.085	0.847
BAA, ASR	3	16	0.754	0.105	0.818

Table 5 The experimental database: total duration and number of programmes per genre

	T.Show	Comm.	Music	Cart.	Foot.	News	Weat.For.
Hours	44.2	3.5	3.9	18.8	17.6	21.2	2
# Programmes	60	67	60	59	22	63	65

Table 4 summarises the best network architectures obtained from our analysis for different concatenations of the selected features[8] and for different number of hidden layers and hidden neurons.

6 Experimental evaluation

6.1 The experimental dataset

The rules for the construction of an experimental database depend on the aim of the research work under investigation. The final *goal* of this work was to classify complete programmes, since the reference domain of our research is a large broadcast archive (RAI). In particular, we tested the ability of our system in classifying television programmes with variable duration into seven genres, named *newscasts, commercials, cartoons, football, music, weather forecasts* and *talk shows*.

In literature it is common the attempt at classifying clips of content instead of semantically coherent units. Though this approach has the evident advantage of making the systems able to detect genres analysing only few minutes of content, we believe that it may produce less useful results when the techniques are applied in contexts where objects are typically classified and accessed as wholes (e.g. digital libraries). Besides, complete programmes are often very complex and shirk from being fully characterised only by a short clip. By classifying a clip, we would be able to declare the local genre of a programme, without being able to say anything about the global content genre, which is the most important aspect in digital libraries search and retrieval systems. Furthermore, clip selection may be prone to a bias effect introduced by whom is selecting the clip boundaries. Complete programmes are editorially controlled by producers and publishers, therefore our potentially biasing mediation is not present. Our experimental database collects thus more than 100 h of complete television programmes from the daily programming of national and regional broadcasters (i.e. three national public channels and nine regional private channels), as detailed in Table 5. Programmes were selected to ensure genre representativeness without loosing in generality. We constructed the set by balancing the number of programmes per genre, respecting the genre frequency observed in the average broadcast schedule of a closed period of time. Programmes were selected randomly in a wider period of time (i.e. between February 2006 and July 2006), to avoid bias effects due to local reuse of material in the short term broadcast (e.g. newscasts and music).

Football is considered as a particular instance of the more general *sports* genre. Talk shows, music programmes and sports programmes are, in turn, kinds of entertainment programmes. Weather forecasts and newscasts are kinds of information

[8] See Table 1 to recall the meaning of the acronyms for the programme surrogate features.

programmes. Finally, commercials are intended as a sub-class of communication programmes. These genres are fairly representative of the programme formats that are currently produced and distributed either through the traditional distribution channels, such as broadcast, cable and satellite, or through new platforms like the Internet or mobile phones.

The newscast sequences were captured from many different news programmes: RAIUNO “TG1” (12 programmes), RAIDUE “TG2” (eight programmes), RAITRE “TG3” (ten programmes), RAITRE “TGR” (15 programmes), RAITRE “TGSport” (four programmes), Telecupole “TG4” (two programmes), Telesubalpina “Il Regionale” (two programmes), Telesubalpina “TG2000” (two programmes), E21 “Tg”, Primantenna “News”, Quartarete “TG4”, Retesette “Informasette”, Telecity “TG7”, Telestudio “Telenews”, Telesubalpina “TG Flash”, Videogruppo “Videonotizie” (one programme for each of the listed newscasts). The news sequences usually have a regular structure, characterised by anchor shots (i.e. repeated and spread over the time presenter shots that characterise structural features) and report shots (i.e. where either the anchorperson or the reporter is talking over an external report). All the news programmes have duration between about 10 min to about 30 min.

The commercials were collected from different channels at different times and days. There are 67 commercial blocks in total. Each block has duration between approximately 1 to 6 min. Each single commercial is separated from the others by some black frames. Commercials are usually characterised by higher volume audio and by the contemporaneous presence of speech and music in the audio stream (i.e. aural features).

The cartoons videos were acquired from a number of kids programmes, including RAIDUE “Classici Disney” (nine programmes), RAIDUE “Random” (11 programmes), RAITRE “La Melevisione” (seven programmes), Quartarete “K2” (16 programmes) and Telestudio “Buonanotte Bambini” (16 programmes). Cartoons programmes are 5 to 50 min long. As cartoons show a great variety of characteristics, depending on e.g. the production date and the age of the target audience, the cartoons genre has probably the most variable semantic content.

Most of the football matches were captured from the “Germany 2006” FIFA World Cup (18 programmes). Other sources are the test match “Switzerland Vs. Italy” (two programmes) and the Italian Soccer Tim Cup final first match “Roma Vs. Inter” (two programmes).

The music genre collects programmes broadcasted by Radio Italia TV, and includes both live concerts and music video clips of Italian singers or groups. Each music video clip includes only one song of duration variable between 2 and 6 min. In addition, some clips were extracted from RAIUNO “Domenica In” (seven programmes with length from 1 to 12 min).

The weather forecast class includes daily regional and national weather forecasts (25 programmes and 40 programmes, respectively), each of duration between 1 and 3 min. Most of the weather forecasts programmes are produced using video showing satellite images and a voice who talks about what is happening in the video (thus related to visual and aural features). Sometimes, there can be the contemporary presence of a presenter (thus related to cognitive features).

The talk shows class includes a broad variety of programmes, including game shows, informative shows, simulated legal encounters, sport shows and candid camera shows.[9] Many talk shows allow members of the public to join in, through telephone calls, letters, e-mails and Internet chat channels. In this work, only in-studio talk shows (i.e. talk shows where one or more hosts discusses face-to-face about e.g. politics, current affairs, religion or other topics with guests) were captured. Each talk show programme has duration between 18 min and 2 h.

6.2 Experimental settings

In the experimental prototype we used the following libraries. The feature extraction library was entirely developed by us in ANSI C++ programming language for UNIX platforms. The face detection task was performed using the RTFaceDetection library[10] offered by Fraunhofer IIS. The speech-to-text task was performed using an engine for the Italian language supplied by ITC-IRST (Centre for Scientific and Technological Research).[11] The neural network classifier was implemented using the Fast Artificial Neural Network (FANN) library.[12]

We set the desired error $\varepsilon = 10^{-4}$ and the maximum number of epochs $M_s = 10^4$. Based on the selection criteria presented in Section 5, the optimal network architecture was set to 210:16:7 (i.e. 210 input neurons, 1 hidden layer with 16 hidden neurons, seven output neurons) for the low-level visual sub-network, 17:64:7 for the structural sub-network, 21:16:7 for the cognitive sub-network and 8:32:7 for the aural sub-network.

Finally, we used K-fold cross validation with $K = 6$, i.e. we partitioned our experimental database $\mathcal{D}$ in six parts $l_1, l_2, ..., l_6$, $\cup_{i=1}^{6} l_i = D$, $l_i \cap l_j = \emptyset\ \forall i \neq j$ and made six distinct training-testing rounds, each selecting a part l_i as the test set $\mathcal{T}$ and the remaining part $\mathcal{L}' = \cup_{j \neq i} l_j$ as training set. The item set subdivision was done taking care of respecting the observed genre frequencies in $\mathcal{D}$.

6.3 Experimental results

This section illustrates the experimental results of the proposed genre classification system. Both the individual effectiveness of each descriptor and the combined performance of the ensemble system are analysed.

Let $\Omega = \{\omega_1, \omega_2, ..., \omega_{N_\omega}\}$ be the set of classes to be discriminated, we calculated the classification accuracy by adding up the correctly classified testing items over all classes:

$$\alpha_t = \frac{1}{N_p} \sum_{i=1}^{N_\omega} |\omega_i^{(right)}|, \tag{12}$$

[9] See http://www.museum.tv/archives/etv/T/htmlT/talkshows/talkshows.htm for an introduction to the history of TV talk shows (Last accessed: September 27th, 2007).

[10] A demo version (valid for 60 days) of the RTFaceDetection library is available for download at: http://www.iis.fraunhofer.de/EN/bf/bv/kognitiv/biom/dd.jsp (last accessed: November 21st, 2007).

[11] http://www.itc.it/irst (last accessed: May 13th, 2008).

[12] http://leenissen.dk/fann (last accessed: May 13th, 2008).

Table 6 Individual classification accuracy for each genre and for each information source (unit: 100%)

	T.Show	Comm.	Music	Cart.	Foot.	News	Weath.For.
F	86.7	89.5	73.3	96.6	100	85.7	87.6
S	91.6	77.6	61.6	88.1	100	90.4	98.4
C	78.3	62.6	45	67.7	59	68.2	89.2
A	73.3	62.7	98	67.8	50	85	87.7

where N_p is the total number of items in the testing set and $|\omega_i^{(right)}|$ is the number of testing items belonging to the class i correctly classified.

6.3.1 Examination of the intrinsic characteristics of TV genres

Genre theory states that objects (e.g. written texts, pieces of music, paintings, TV programmes) sharing a common purpose tend to share common characteristics. In order to identify by which properties a certain TV genre is distinguishable from the others, we evaluated the classification accuracy achieved considering the best feature combination according to the information source from which it was extracted, e.g. the vector ($\mathbf{H}$, $\mathbf{S_{at}}$, $\mathbf{V}$, $\mathbf{Y}$, $\mathbf{C_{on}}$, $\mathbf{D_{ir}}$, $\mathbf{T_{em}}$) for the low-level visual features component **F**. Based on what aggregations contributed most to the classification accuracy, a certain genre can be then decomposed into its main intrinsic characteristics (its *multimodal essence*). Table 6 explains this concept, where the element (i, j) indicates the accuracy with which a classifier based only on information source of type i explains the

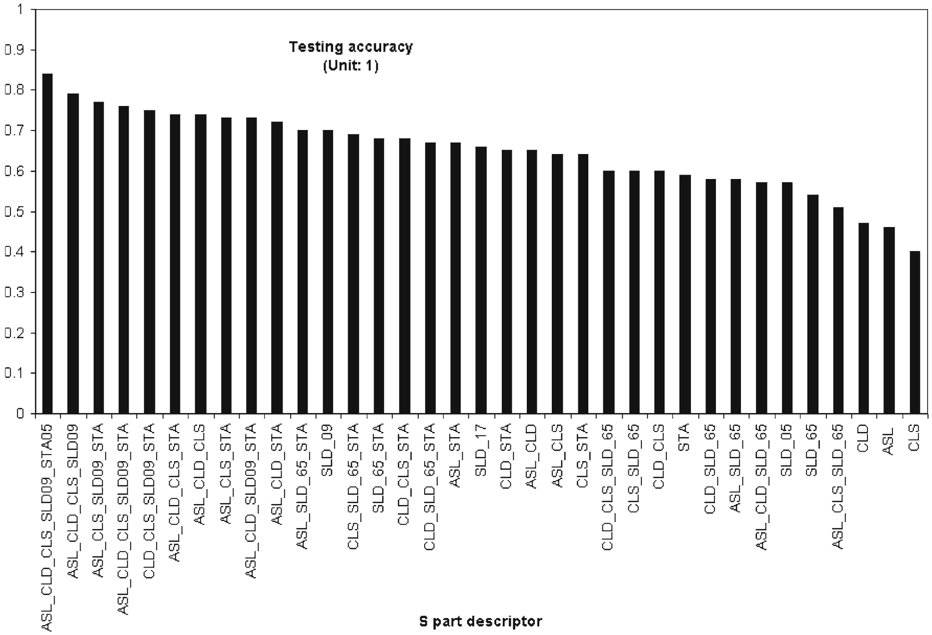


Fig. 8 Individual performances of the structural descriptors

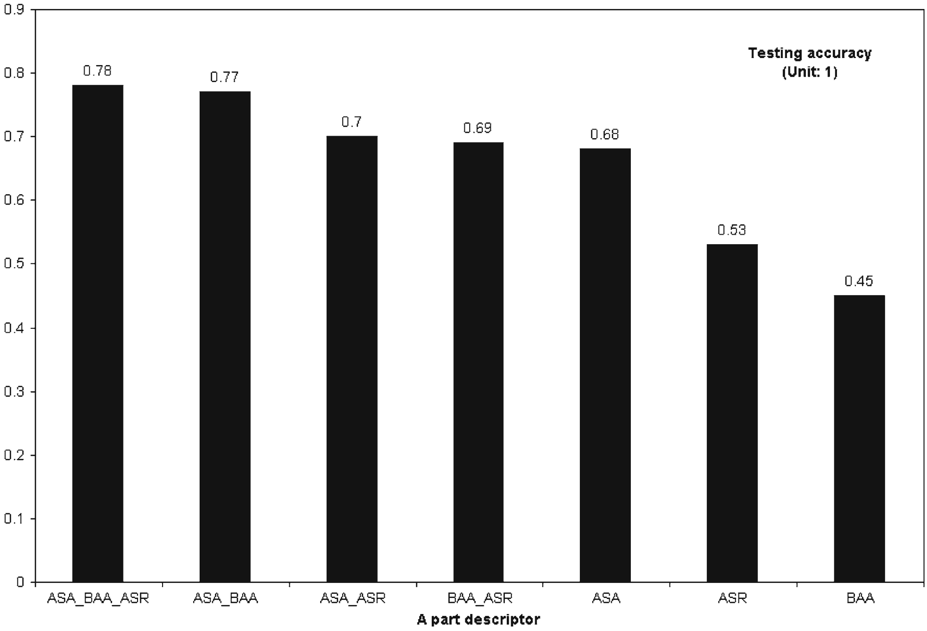


Fig. 9 Individual performances of the aural descriptors

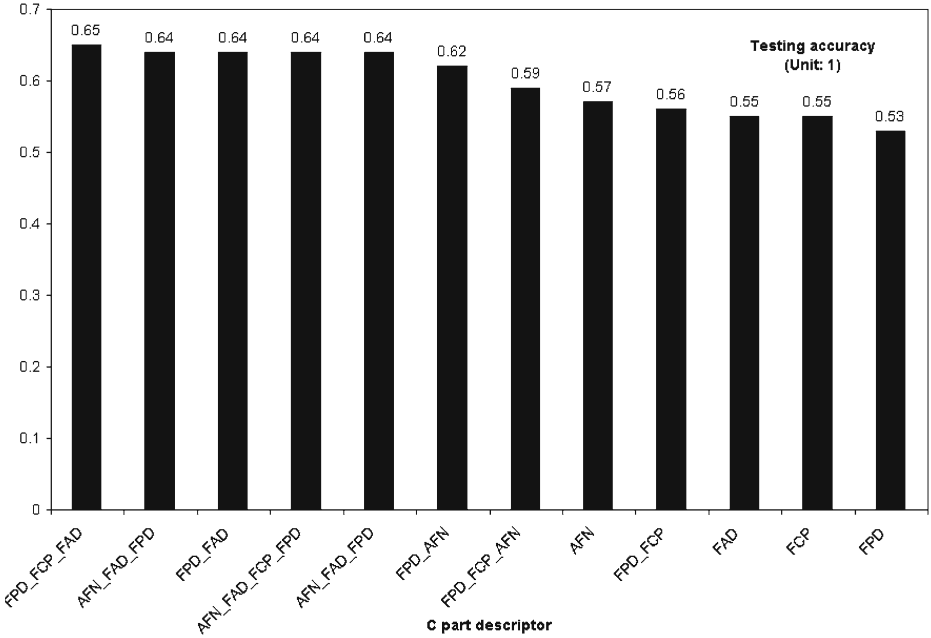


Fig. 10 Individual performances of the cognitive descriptors

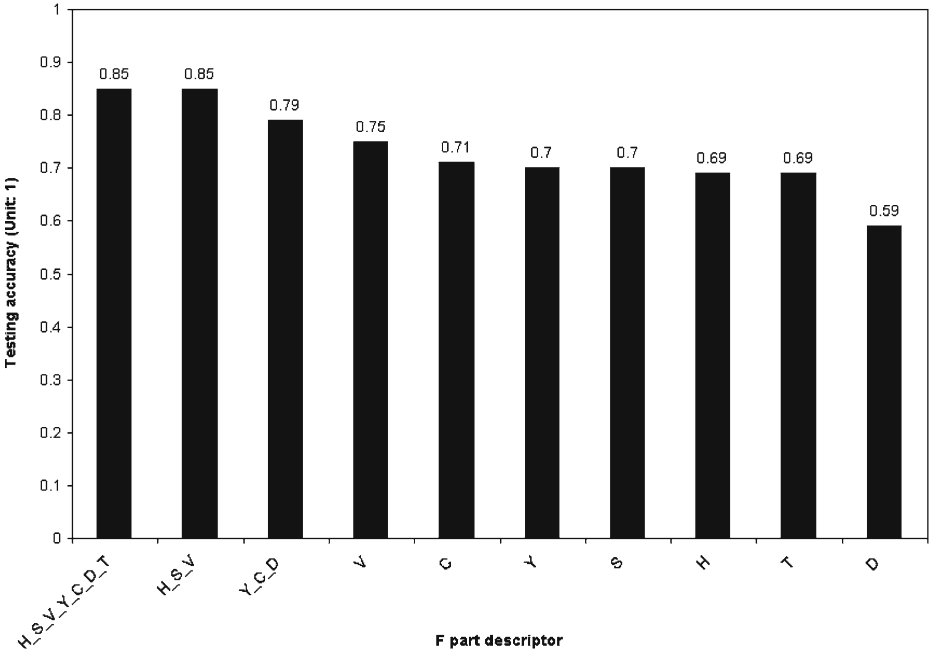


Fig. 11 Individual performances of the visual, perceptual descriptors

genre *j*. Results show that e.g. football matches have perceptual (**F**) and structural (**S**) distinctive nature, while music clips mostly acoustic-related (**A**) distinctive features.

6.3.2 Performance of individual descriptors

In order to evaluate the individual performance of each feature we calculated the classification accuracy (α_t) for each part of the programme surrogate. Figures 8, 9, 10 and 11 show the results of our experiments for proving the power of the individual descriptors by themselves. The classification accuracy is plotted with respect to different concatenations of the programme surrogate sub-parts. Here, a small increasing of the testing accuracy when adding a descriptor indicates a lower discrimination power of the descriptor itself. For example, it is shown that there is not a significant improvement in the performance of the classification system if we consider the concatenation of colour and texture features instead of considering only colour features (see Fig. 11). However, none of the considered information sources is sufficient alone to satisfactorily distinguish genres each other.

Table 7 Ensemble classification accuracy (unit: 100%)

Part	A C F S	A F S	F	S	A	C
Mean	94.9	94.1	85.6	83.3	78.7	65.2
S.Dev.	1.12	1.07	1.64	1.27	1.83	1.76

Table 8 Video genre classification accuracy (unit: 100%)

Exp_1	Exp_2	Exp_3	Exp_4	Exp_5	Exp_6	Average
95.5	93.5	95.5	95.6	95.5	93.8	**94.9**

6.3.3 Performance of combined descriptors

Table 7 reports the classification accuracy for some ensemble combinations, averaged on the six experiments. Optimal performance (i.e. the ensemble with the lowest standard deviation of the testing accuracy) was achieved with the ensemble of the aural, structural and low-level visual parts of the programme surrogate, while the best performance (i.e. the ensemble with the highest mean value of the testing accuracy) was achieved considering altogether the parts of the programme surrogate ($\sim 95\%$). However, it can be noticed that the use of the cognitive part of the surrogate does not increase significantly the testing accuracy w.r.t. the use of only the other parts of the surrogates.

Table 8 shows the classification accuracy for the six experiments, achieved selecting all the extracted features of the programme surrogate.

Table 9 shows the confusion matrix averaged on the six experiments, where the element in position (i, j) is the percentage of programmes actually belonging to the genre i and classified as the genre j. For example, it is shown that the 98.5% of weather forecast programmes were correctly classified as weather forecasts, while the remaining 1.5% was wrongly classified as music content. As expected, some *news* and *talk shows* tend to be confused each other, due to their common cognitive and structural properties. For some genres the overall accuracy is ~100% (i.e. *football, cartoons* and *weather forecasts*). This demonstrates that: (1) these genres are well recognisable and separated from the other ones; and (2) the extracted features are effective when applied to the examined task.

Finally, in order to improve the performance of our ensemble classification system, we weighted the outputs of each individual information source according to results from Section 6.3.1. Thus, the outputs of each sub-network are now tuned according to the following equation:

$$\Phi^{(p,n)} = (\beta_1^{(n)}\phi_1^{(p,n)}, \ldots, \beta_{N_\omega}^{(n)}\phi_{N_\omega}^{(p,n)}), \;\; n = 1, \ldots, 4 \tag{13}$$

where:

- p is a TV programme to be classified;
- $\Omega = \{\omega_1, \ldots, \omega_{N_\omega}\}$ is the set of available genres

Table 9 Video genre confusion matrix (unit: 100%)

	T.Show	Comm.	Music	Cart.	Foot.	News	Weat.For.
T.Show	**90**	0	0	0	0	10	0
Comm.	0	**97**	1.5	0	0	1.5	0
Music	1.7	5	**88.2**	1.7	0	1.7	1.7
Cart.	0	0	0	**100**	0	0	0
Foot.	0	0	0	0	**100**	0	0
News	4.8	1.6	0	0	0	**93.6**	0
Weat.For.	0	0	1.5	0	0	0	**98.5**

Table 10 Per class classification accuracy for unweighted and weighted network outputs (Unit: 100%)

	T.Sh.	Com.	Mus.	Cart.	Foot.	News	Weat.F.	Mean
Unweighted	90	97	88	100	100	94	98	95
Weighted	93	95	93	100	100	91	98	96

- $\beta_i^{(n)}$ $(i = 1, \ldots, N_\omega)$ is a factor that takes into account how well the information source n produces classification of genre i (e.g. from Table 6, $\beta = 0.867$ for talk shows when considering the **F** part of the programme surrogate);
- $\phi_i^{(p,n)}$ is the output of the ith neuron of the sub-network n.

Table 10 compares the cases of unweighted and weighted network outputs. It shows that, since different feature set has different influence on video genre classification, the overall classification accuracy increases when this influence is considered in the classification model [30].

7 Conclusions

7.1 Examples of end-user applications

This section outlines some examples of practical applications that involve automatic genre classification. According to [44] genre classification is a fundamental step for building a semantic index of video contents.

The use of automatic genre recognition systems would bring benefits not only to information and entertainment industries but also to end-users. In fact, on one hand, TV broadcasters would be able to manage more efficiently their multimedia archives, enabling user oriented, quick and advanced video retrieval facilities, thus decreasing costs for production, editing and fruition. On the other hand, end-users would be allowed to create personalised TV programme lists using video-on-demand (VoD) and interactive TV services.

A second example regards TV commercials detection. This topic mainly attracts consumer electronics companies that aim at developing an automatic way for their digital video recorders to detect and remove commercials from recorded programmes. Moreover, also public institutions might be interested in TV commercials detection. For example, they could be interested in monitoring television commercials for accuracy and compliance with antitrust policies.

A last, but not least, example concerns the possibility of analysing and annotating uncontrolled multimedia contents (e.g. user-generated videos) according to standardised criteria.

7.2 Final comments and future work

In this paper we presented a set of features for the multimedia genre recognition task, and a genre analysis and classification system based on those features. Some of these features are well defined in literature (e.g. the average shot length and the average speech rate), other ones are new (e.g. the shot cluster duration and saturation, the face covering percentage and the shot temporal activity).

Due to the heterogeneous nature of the characteristics of multimedia contents, in our system features are first grouped into subsets, each representing a particular aspect of multimedia data. Each subset is then treated as the input vector for a neural network classifier. The outputs of all individual classifiers are combined by an ensemble method to get the final classification.

We extensively tested our system in the domain of complete TV programmes, and using seven genres. The obtained accuracy in the classification task is outstanding and demonstrates the effectiveness of the selected features, due to their capability of capturing independent aspects of the examined genres. Another result of our research consists in the analysis of the correlation between the nature of the information sources and their discrimination power w.r.t. genres.

With our current software version, which has still great room for optimisation, it took about 50 h for completing the feature extraction and the training tasks on approximately 110 h of video material, on a dual Xeon at 2.8 GHz clock, i.e. approximately 1:2. However, the high level of parallelisation of the deployed architecture (see Fig. 1) allows for substantial time reduction provided by the insertion of additional computing nodes. This aspect will be one of the main future works from the implementation perspective.

Other future work will investigate the development of new features, such as the mutual relationships between shot clusters (e.g. adjacency, concatenation, inclusion) or other text-based descriptors, such as lists of keywords (e.g. nouns, verbs and places) from the transcribed text. In addition, the introduction of new genres and sub-genres in the reference taxonomy will be also considered.

A final consideration concerns the multiformity of multimedia objects, which are typically hardly assignable to a single genre. In fact, in real world scenarios, a single object can belong to none, one or more genres at the same time (e.g. *entertainment* programmes may be usually associated to several kinds of different programme formats). In addition, the boundary between genres is not sharp. To overcome these limitations, we are exploring the use of fuzzy classifiers, which capture additional information about the certainty degree of the classification decision, to model both genres fuzziness and multimedia objects multiformity w.r.t. genres. Our preliminary experimental results demonstrated the potential of this approach [25, 26].

References

1. Albiol A, Fullá MJCh, Albiol A, Torres L (2004) Commercials detection using HMMs. In: International workshop on image analysis for multimedia interactive services. Lisboa, Portugal
2. Bellman R (1961) Adaptive control processes: a guided tour. Princeton Univ. Press
3. Blum DW (1992) Method and apparatus for identifying and eliminating specific material from video signals. US Patent no. 5151788
4. Boggs J, Petrie DW (2006) The art of watching films with tutorial CD-ROM. McGraw-Hill
5. Brugnara F, Cettolo M, Federico M, Giuliani D (2000) A system for the segmentation and transcription of italian radio news. In: RIAO, content-based multimedia information access. Paris, France
6. Ćalić J (2004) Highly efficient low-level feature extraction for video representation and retrieval. PhD thesis, University of London
7. Chellappa R, Wilson CL, Sirohey S (1995) Human and machine recognition of faces: a survey. Proc IEEE 83(5):705–740 (May)

8. Cheng W, Liu C, Wang X (2006) A rough set approach to video genre classification. In: 8th international conference on advanced concepts for intelligent vision systems (ACIVS'06). Antwerp, Belgium, pp 1210–1220 (September)
9. Covell M, Baluja S, Fink M (2006) Advertisement detection and replacement using acoustic and visual repetition. In: IEEE 8th workshop on multimedia signal processing (MMSP2006). Victoria, BC, pp 461–466 (October)
10. Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13(1): 21–27
11. Dimitrova N, Agnihotri L, Wei G (2000) Video classification based on HMM using text and faces. In: European conference on signal processing. Tampere, Finland
12. Dimitrova N, Jeannin S, Nesvadba J, McGee T, Agnihotri L, Mekenkam G (2002) Real time commercial detection using MPEG features. In: Proc. 9th int. conf. on information processing and management of uncertainty in knowledge-based systems (IPMU 2002). Annecy, France, pp 481–486 (Invited paper)
13. Dinh PQ, Dorai C, Venkatesh S (2002) Video genre categorization using audio wavelet coefficients. In: ACCV2002: the 5th Asian conference on computer vision. Melbourne, Australia (January)
14. Dorado A, Calic J, Izquierdo E (2004) A rule-based video annotation system. IEEE Trans Circuits Syst Video Technol 14(5):622–633
15. EBU-UER (2007) Escort 2007. Technical Review 3322, EBU
16. Fischer S, Lienhart R, Effelsberg W (1995) Automatic recognition of film genres. In: ACM multimedia 1995. San Francisco, CA, pp 295–304 (November)
17. Glasberg R, Samour A, Elazouzi K, Sikora T (2005) Cartoon-recognition using video & audio-descriptors. In: 13th European signal processing conference (EUSIPCO2005). Antalya, Turkey (September)
18. Goh KS, Miyahara K, Radhakrishan R, Xiong Z, Divakaran A (2004) Audio-visual event detection based on mining of semantic audio-visual labels. Technical Report 2004-008, Mitsubishi Electric Research Laboratory (MERL)
19. Ianeva TI, de Vries AP, Rohrig H (2003) Detecting cartoons: a case study in automatic video-genre classification. In: IEEE international conference on multimedia and expo (ICME'03), pp 449–452 (July)
20. Igel C, Hüsken M (2000) Improving the Rprop learning algorithm. In: Proceedings of the second international symposium on neural computation, NC2000
21. Jolliffe IT (2002) Principal component analysis. Springer
22. Liu Z, Huang J, Wang Y (1998) Classification of TV programs based on audio information using hidden Markov model. In: IEEE 2nd workshop on multimedia signal processing (MMSP '98). Redonda Beach, CA, USA, pp 27–32 (December)
23. Liu Z, Huang J, Wang Y, Chen T (1997) Audio feature extraction and analysis for scene classification. In: IEEE workshop on multimedia signal processing (MMSP'97), pp 343–348
24. Lo Iacono A, Colamussi M (2005) Rai click—"I want my own TV". Technical Review 303, EBU (July)
25. Messina A, Montagnuolo M (2008) Fuzzy mining of multimedia genre applied to television archives. In: IEEE international conference on multimedia and expo. Hannover, Germany, 23–26 June 2008
26. Messina A, Montagnuolo M (2008) Multimedia genre characterisation with fuzzy embedding classifiers. In: International workshop on ambient media delivery and interactive television (AMDIT2008). Quebec City, Canada (February)
27. Messina A, Montagnuolo M, Sapino ML (2006) Characterizing multimedia objects through multimodal content analysis and fuzzy fingerprints. In: IEEE international conference on signal-image technology and internet-based systems (SITIS'06). Hammamet, Tunisia (December)
28. Montagnuolo M, Messina A (2007) Automatic genre classification of TV programmes using Gaussian mixture models and neural networks. In: DEXA workshops. Regensurg, Germany, pp 99–103 (September)
29. Montagnuolo M, Messina A (2007) Multimedia knowledge representation for automatic annotation of broadcast TV archives. J Digit Inf Manag 5(2):67–74
30. Montagnuolo M, Messina A (2008) Multimodal genre analysis applied to digital television archives. In: Second international workshop on multimedia data mining and management (DEXA-MDMM'08). Turin, Italy, 2 September 2008
31. Novak AP (1988) Method and system for editing unwanted program material from broadcast signals. US Patent no. 4750213

32. Parnal S, Pizzi S (2003) TV anytime: a new standard. EBU diffusion online, 2003/33, August
33. Poli JP, Carrive J (2006) Improving program guides for reducing tv stream structuring problem to a simple alignment problem. In: CIMCA '06: proceedings of the international conference on computational inteligence for modelling control and automation and international conference on intelligent agents web technologies and international commerce, p 31
34. Polikar R (2006) Ensemble based systems in decision making. IEEE Circuits Syst Mag 6(3):21–45
35. Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc
36. Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77(2):257–286
37. Roach M, Mason JS, Pawlewski M (2001) Motion-based classification of cartoons. In: IEEE international symposium on intelligent multimedia, video and speech processing (ISIMP2001), pp 146–149
38. Roach MJ (2002) Video genre classification. PhD thesis, University of Wales Swansea
39. Roach MJ, Mason JSD, Pawlewski M (2001) Video genre classification using dynamics. In: IEEE international conference on acoustics, speech, and signal processing (ICASSP'01), pp 1557–1560
40. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning internal representations by error propagation. In: Parallel distributed processing: volume 1: foundations. The MIT Press, pp 318–362
41. Safavian SR, Landgrebe DA (1991) A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern 21(3):660–674
42. Sánchez JM, Binefa X, Vitriá J, Radeva P (1999) Local color analysis for scene break detection applied to TV commercials recognition. In: VISUAL '99: proceedings of the third international conference on visual information and information systems, pp 237–244
43. Satterwhite B, Marques O (2004) Automatic detection of television commercials. IEEE Potentials 23(2):9–12
44. Snoek C, Worring M (2005) Multimodal video indexing: a review of the state-of-the-art. Multimedia Tools and Applications 25(1):5–35
45. Swain MJ, Ballard DH (1991) Color indexing. Int J Comput Vis 7(1):11–32 (November)
46. Takagi S, Hattori S, Yokoyama K, Kodate A, Tominaga H (2003) Sports video categorizing method using camera motion parameters. In: IEEE 2003 international conference on multimedia and expo (ICME'03), pp 461–464 (July)
47. Tamura H, Mori S, Yamawaki T (1978) Texture features corresponding to visual perception. IEEE Trans Syst Man Cybern 8(6):460–473
48. Taskiran CM, Delp EJ (2001) Distribution of shot lengths for video analysis. In: Proceedings of SPIE, vol. 4676, pp 276–284
49. Taskiran CM, Pollak I, Bouman CA, Delp EJ (2003) Stochastic models of video structure for program genre detection. In: 8th international workshop on visual content processing and representation (VLBV 2003). Madrid, Spain, pp 84–92 (September)
50. Tekalp M (1995) Digital video processing. Prentice Hall
51. Tomasi C (2005) Estimating Gaussian mixture densities with EM—a tutorial. Technical report, Duke University
52. Truong BT, Venkatesh S, Dorai C (2000) Automatic genre identification for content-based video categorization. In: IEEE 15th international conference on pattern recognition (ICPP'00). IEEE Computer Society, pp 230–233
53. Vakkalanka S, Mohan CK, Kumaraswamy R, Yegnanarayana B (2005) Combining multiple evidence for video classification. In: IEEE international conference on intelligent sensing and information processing (ICISIP'05), pp 187–192 (January)
54. Vapnik VN (1999) The nature of statistical learning theory. Springer
55. Vasconcelos N, Lippman A (2000) Statistical models of video structure for content analysis and characterization. IEEE Trans Image Process 9(1):3–19
56. Vroomen JHM, Collier R, Mozziconacci S (1993) Duration and intonation in emotional speech. In: Eurospeech 1993, pp 577–580
57. Wang J, Xu C, Chang E (2006) Automatic sports video genre classification using pseudo-2D-HMM. In: IEEE 18th international conference on pattern recognition (ICPR'06), pp 778–781
58. Wickenberg-Bolin U, Göransson H, Fryknäs M, Gustafsson MG, Isaksson A (2006) Improved variance estimation of classification performance via reduction of bias caused by small sample size. BMC Bioinformatics 7:127
59. Xu LQ, Li Y (2003) Video classification using spatial-temporal features and PCA. In: IEEE international conference on multimedia and expo (ICME'03), pp 485–488 (July)

60. Yuan X, Lai W, Mei T, Hua XS, Wu XQ, Li S (2006) Automatic video genre categorization using hierarchical SVM. In: IEEE international conference on image processing (ICIP'06). Atlanta, GA, pp 2905–2908 (October)
61. Yuan Y, Song QB, Shen JY (2002) Automatic video classification using decision tree method. In: IEEE 1st international conference on machine learning and cybernetics, vol. 3. Beijing, pp 1153–1157
62. Zhiwen Y, Xingshe Z, Jianhua G, Zhiyi Y (2004) Fuzzy clustering for tv program classification. In: IEEE international conference on information technology: coding and computing (ICIT'04), pp 658–662 (April)

Maurizio Montagnuolo Born in 1975, Maurizio Montagnuolo received his Laurea degree in Telecommunications Engineering from the Polytechnic of Turin in 2004, after developing his thesis at the RAI Research Centre. Currently, he is attending the Ph.D. course in "Business and Management" at the University of Turin, in collaboration with RAI, and supported by EuriX S.r.l., Turin. His main research interests concern the semantic classification of audiovisual content.

Alberto Messina is from the RAI—Radiotelevisione Italiana Centre for Research and Technological Innovation (CRIT), Turin.

He began his collaboration as a research engineer with RAI in 1996, when he completed his MS Thesis in Electronic Engineering (at Politecnico di Torino) about objective quality evaluation of MPEG2 video coding. After starting his career as a designer of RAI's Multimedia Catalogue, he has been involved in several internal and international research projects in the field of digital archiving, with particular emphasis on automated documentation, and automated production. His current interests are ranging from file formats and metadata standards to the domain of content analysis and information extraction algorithms, where he now concentrates his main focus. Recently, he has started promising research activities concerning semantic information extraction from the numerical analysis of audiovisual material, particularly in the field of conceptual characterisation of multimedia objects, genre classification of multimedia items, automatic editorial segmentation of TV programmes. He is also author of technical and scientific publications in this subject area. He has extensive collaborations with the local University of Torino—Computer Science Department, which include common research projects and students' tutorship. To complete his scientific formation, he has recently decided to take a PhD in the area of Computer Science.

He is active member of several EBU projects including P/TVFILE, P/MAG and P/CP, chairman of the P/SCAIE project dealing with automatic metadata extraction techniques. He is currently working in the EU PrestoSpace project in the Metadata Access and Delivery area. He has served as Programme Committee Member in a Special Track of the 10th Conference of Italian Association of Artificial Intelligence, and in the First Workshop on Ambient media Delivery and Interactive Television (AMDIT08).