TV Genre Classification Using Multimodal Information and Multilayer Perceptrons

Maurizio Montagnuolo[1,⋆] and Alberto Messina[2]

[1] Università degli Studi di Torino, Dip. di Informatica, Torino, Italy
montagnuolo@di.unito.it
[2] RAI Centro Ricerche e Innovazione Tecnologica, Torino, Italy
a.messina@rai.it

Abstract. Multimedia content annotation is a key issue in the current convergence of audiovisual entertainment and information media. In this context, automatic genre classification (AGC) provides a simple and effective solution to describe video contents in a structured and well understandable way. In this paper a method for classifying the genre of TV broadcasted programmes is presented. In our approach, we consider four groups of features, which include both low-level visual descriptors and higher level semantic information. For each type of these features we derive a characteristic vector and use it as input data of a multilayer perceptron (MLP). Then, we use a linear combination of the outputs of the four MLPs to perform genre classification of TV programmes. The experimental results on more than 100 hours of broadcasted material showed the effectiveness of our approach, achieving a classification accuracy of $\sim 92\%$.

1 Introduction

Improvements in video compression, in conjunction with the availability of high capacity storage devices have made possible the production and distribution to users of digital multimedia content in a massive way. As large-scale multimedia collections come into view, efficient and cost-effective solutions for managing these vast amounts of data are needed. Current information and communication technologies provide the infrastructure to transport bits anywhere, but on the other hand, our current ability on user-oriented multimedia classification is not still mature enough, due to the lack of good automated semantic extraction and interpretation algorithms. The problem concerning the reduction of the "semantic gap" (i.e. combining and mapping low-level descriptors automatically extracted by machines to high-level concepts understandable by humans) is the main challenge of the Video Information Retrieval (VIR) research community. A critical review of the applicability of VIR technologies in real industrial scenarios in presented in [14].

In the television programme area, the classification of video content into different genres (e.g. documentary, sports, commercials, etc.) is an important topic

⋆ PhD student, supported by EuriX S.r.l., Torino, Italy. – www.eurixgroup.com

R. Basili and M.T. Pazienza (Eds.): AI*IA 2007, LNAI 4733, pp. 730–741, 2007.

of VIR. Even if the definition of genre may depend on social, historical, cultural and subjective aspects, thus resulting in fuzzy boundaries between different genres, many common features characterise objects belonging to the same genre. Therefore, automatic genre classification provides a good way to capture semantic information about multimedia objects. For instance, in video production, genre is usually intended as a description of what type of content TV viewers expect to watch.

In this work, a solution for automatic genre classification is presented. In particular, we describe a framework that is able to discern between TV commercials, newscasts, weather forecasts, talk shows, music video clips, animated cartoons and football match videos. 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 present approach is based on two foundations: (i) *Multimodal content analysis* to derive a compact numerical representation (here in after called *pattern vector* – PV) of the multimedia content; and (ii) *Neural Network training process* to produce a classification model from those pattern vectors. Neural networks are successfully used in pattern recognition and machine learning [21]. Our system uses four multilayer perceptrons to model both low-level and higher level properties of multimedia contents. The outputs of these MLPs are combined together to perform genre classification.

The remaining of the paper is organised as follows. The sets of extracted features are detailed in Section 2. The process of genre classification based on neural networks is described in Section 3. Experimental results are given in Section 4. Finally, conclusions and future work are outlined in Section 5.

2 Proposed Feature Sets

Multimodal information retrieval techniques combine audio, video and textual information to produce effective representations of multimedia contents [22]. Our system is multimodal in that it uses a pattern 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, structural-syntactic information and cognitive information. In the broadcast domain in which we operate, modality information concerns the physical properties of audiovisual content, as they can be perceived by users (e.g. colours, shapes, motion). Structural-syntactic information describes spatial-temporal layouts of programmes (e.g. relationships between frames, shots and scenes). Cognitive information is related to high-level semantic concepts inferable from the fruition of audiovisual content (e.g. genre, events, faces). An exhaustive analysis concerning the representation of multimedia information content of audiovisual material can be found in [16].

Starting from the basic media types introduced above, we derive the TV programme pattern vector $\boldsymbol{PV} = (\boldsymbol{V_c}, \boldsymbol{S}, \boldsymbol{C}, \boldsymbol{A})$. The pattern vector collects four

sets of features that capture visual ($\boldsymbol{V_c}$), structural ($\boldsymbol{S}$), cognitive ($\boldsymbol{C}$) and aural ($\boldsymbol{A}$) properties of the video content. We have originally designed some of these features to reflect the criteria used by editors in the multimedia production process. For example, commercials are usually characterised by a rapid mix of both music and speech. On the other hand, the majority of talk shows present lengthy shots and have less music and more speech contribution. The physical architecture designed and implemented to calculate the pattern vector is presented in [15]. The four feature sets included in the pattern vector are detailed in the following subsections.

2.1 The Low-Level Visual Pattern Vector Component

The low-level visual pattern vector component includes seven features. Colours are represented by hue (H), saturation (S) and value (V) [23]. Luminance (Y) is represented in a grey scale in the range [16, 233], with black corresponding to the minimum value and white corresponding to the maximum value. Textures are described by contrast (C) and directionality (D) Tamura's features [24]. Temporal activity information (T) is based on the displaced frame difference (DFD) [26] for window size $t = 1$.

First, for each feature we compute a 65-bin histogram, where the last bin collects the number of pixels for which the computed value of the feature is undefined. The low-level visual features are originally computed on a frame by frame basis (e.g. there is one hue histogram for each frame). To provide a global characterisation of a TV programme, we use cumulative distributions of features over the number of frames within the programme. Then, we model each histogram by a 10-component Gaussian mixture, where each component is a Gaussian distribution represented by three parameters: weight, mean and standard deviation [27]. Finally, we concatenate these mixtures to obtain a 210-dimensional feature vector $\boldsymbol{V_c} = (\boldsymbol{H}, \boldsymbol{S}, \boldsymbol{V}, \boldsymbol{Y}, \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{T})$.

2.2 The Structural Pattern Vector Component

The structural component of the pattern vector is constructed from the structural information extracted by a shot detection module. First, a TV programme is automatically segmented into camera shots. We define a shot as a sequence of contiguous frames characterised by similar visual properties. Then, we derive two features S_1 and $\boldsymbol{S_2}$, obtaining a 66-dimensional feature vector $\boldsymbol{S} = (S_1, \boldsymbol{S_2})$. S_1 captures information about the rhythm of the video:

$$S_1 = \frac{1}{F_r N_s} \sum_{i=1}^{N_s} \Delta s_i \tag{1}$$

where F_r is the frame rate of the video (i.e. 25 frames per second), N_s is the total number of shots in the video and Δs_i is the shot length, measured as the number of frames within the i^{th} shot.

$\boldsymbol{S_2}$ describes how shot lengths are distributed along the video. $\boldsymbol{S_2}$ is represented by a 65-bin histogram. Bins 0 to 63 are uniformly distributed in the range [0, 30s], and the 64^{th} bin contains the number of shots whose length is greater than 30 seconds. All histograms are normalised by N_s, so that their area sums to one.

2.3 The Cognitive Pattern Vector Component

The cognitive component of the pattern vector is built by applying face detection techniques [5]. We use this information to derive the following three features:

$$C_1 = \frac{N_f}{D_p} \tag{2}$$

where N_f is the total number of faces detected in the video and D_p is the total number of frames within the video.

The second feature ($\boldsymbol{C_2}$) describes how faces are distributed along the video. $\boldsymbol{C_2}$ is expressed by a 11-bin histogram. The i^{th} $(i = 0, \ldots, 9)$ bin contains the number of frames that contain i faces. The 11^{th} bin contains the number of frames that depict 10 or more faces.

The last feature ($\boldsymbol{C_3}$) describes how faces are positioned along the video. $\boldsymbol{C_3}$ is represented by a 9-bin histogram, where the i^{th} bin represents the total number of faces in the i^{th} position in the frame. Frame positions are defined in the following order: *top-left, top-right, bottom-left, bottom-right, left, left, top, bottom, centre.*

All histograms are normalised by N_f, so that their area sums to one. We concatenate C_1, $\boldsymbol{C_2}$ and $\boldsymbol{C_3}$, to produce a single 21-dimensional feature vector $\boldsymbol{C} = (C_1, \boldsymbol{C_2}, \boldsymbol{C_3})$.

2.4 The Aural Pattern Vector Component

The aural component of the pattern vector is derived by the audio analysis of a TV programme. We segment the audio signal into seven classes: *speech, silence, noise, music, pure speaker, speaker plus noise, speaker plus music.* These computed duration values, normalised by the total duration of the video, are stored in the feature vector $\boldsymbol{A_1}$. In addition, we use a speech-to-text engine [2] to produce transcriptions of the speech content and to compute the average speech rate in the video (A_2). The aural component thus results in a 8-dimensional feature vector $\boldsymbol{A} = (\boldsymbol{A_1}, A_2)$.

3 Video Genre Classification

The process of video genre classification is shown in Figure 1. Let p be a TV programme to be classified and $\Omega = \{\omega_1, \omega_2, \ldots, \omega_{N_\omega}\}$ be the set of available genres. We firstly derive the pattern vector of p as described in Section 2. Each

part of the pattern vector is the input of a neural network. We trained all networks by the iRPROP training algorithm described in [9]. The training process runs until the desired error ε is reached, or until the maximum number of steps MAX_S is exceeded.

In order to define the most suitable network architecture for each part of the pattern vector (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 by Equation 3. The training efficiency combines the training *accuracy* (the ratio of correct items to the total number of items in the training set) and the training *quality* (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].

$$\eta = \frac{2 \cdot accuracy \cdot (1 - quality)}{accuracy + (1 - quality)} \tag{3}$$

   Figure 2 to Figure 5 show the training efficiency for each part of the pattern vector;
2. The *total number of hidden neurons* (HNs) and the *total number of hidden layers* (HLs).

For each NN we have an output vector $\boldsymbol{\Phi}^{(p,n)} = \{\phi_1^{(p,n)}, \ldots, \phi_{N_\omega}^{(p,n)}\}, n = 1, \ldots, 4$ whose element $\phi_i^{(p,n)}$ $(i = 1, \ldots, N_\omega)$ can be interpreted as the membership value of p to the genre i, according to the pattern vector part n. We then combine these outputs to produce an ensamble classifier [17], resulting in a vector $\boldsymbol{\Phi}^{(p)} = \{\phi_1^{(p)}, \ldots, \phi_{N_\omega}^{(p)}\}$, where:

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

We finally select the genre j corresponding to the maximum element of $\boldsymbol{\Phi}^{(p)}$ as the genre with which to classify p.

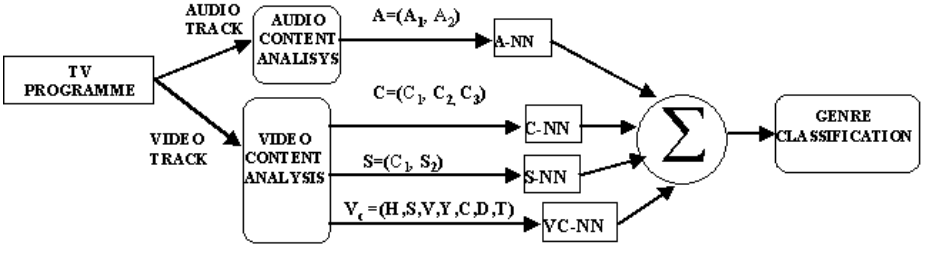

Fig. 1. The video genre classification process

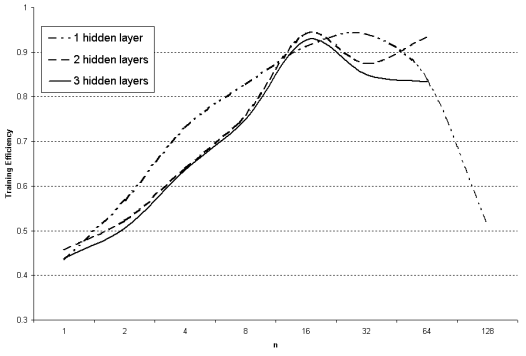

Fig. 2. Training efficiency versus the number of neurons per hidden layer (n) for the aural pattern vector component

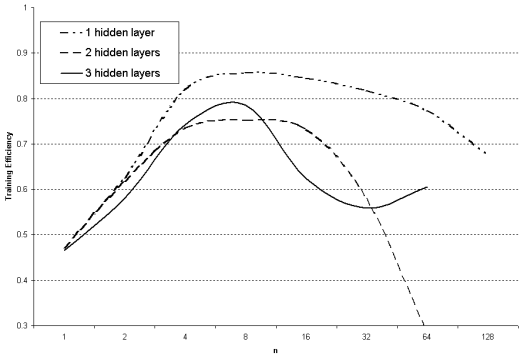

Fig. 3. Training efficiency versus the number of neurons per hidden layer (n) for the structural pattern vector component

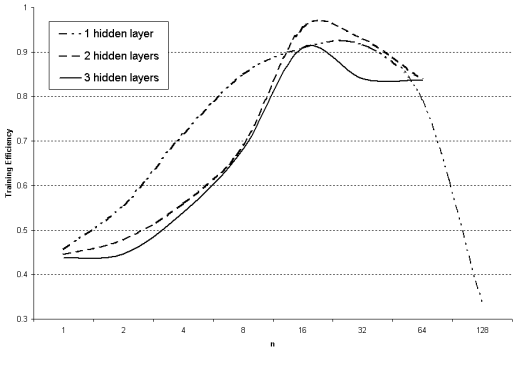

Fig. 4. Training efficiency versus the number of neurons per hidden layer (n) for the cognitive pattern vector component

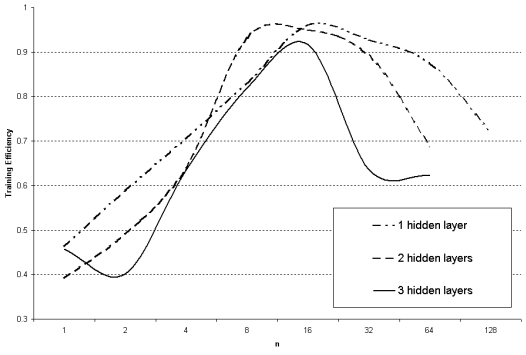

Fig. 5. Training efficiency versus the number of neurons per hidden layer (n) for the visual pattern vector component

4 Experimental Results

4.1 The Experimental Dataset

The experimental dataset collects about 110 hours of complete TV programmes from the daily programming of public and private broadcasters. We built the dataset so that to obtain seven uniformly (w.r.t. the number of occurrences of each genre in the TV programme schedules within a finite period of time) distributed genres: news, commercials, cartoons, football, music, weather forecasts and talk shows. This experimental dataset could be made available for use by researchers. Each TV programme was then manually pre-annotated as belonging to one of the previous genres. Finally, we randomly split the dataset into $K = 6$ disjointed and uniformly distributed (w.r.t. the occurrence of the seven genres in each sub-set) subsets of approximately equal size and used K-fold validation of data. Each subset was thus used once as test set and five times as training set, leading a more realistic estimation of classification accuracy.

4.2 Experimental Settings

In the experimental prototype we used the following libraries:

- The face detection task is performed using the RTFaceDetection library offered by Fraunhofer IIS.[1]
- The speech-to-text task is performed using an engine for the Italian language supplied by ITC-IRST (Centre for Scientific and Technological Research).[2]
- The neural network classifier is implemented using the Fast Artificial Neural Network (FANN) library.[3]

[1] http://www.iis.fraunhofer.de/bv/biometrie/tech/index.html
[2] http://www.itc.it/irst
[3] http://leenissen.dk/fann

Based on the selection criteria presented in Section 3, we chose the following network architectures:

- All networks have one hidden layer and seven output neurons with sigmoid activation functions, whose outputs are in the range [0,1];
- All hidden neurons have symmetric sigmoid activation functions, whose outputs are in the range [-1,1];
- The Aural Neural Network (A-NN) has 8 input neurons and 32 hidden neurons;
- The Cognitive Neural Network (C-NN) has 21 input neurons and 32 hidden neurons;
- The Structural Neural Network (S-NN) has 65 input neurons and 8 hidden neurons;
- The Visual Neural Network (VC-NN) has 210 input neurons and 16 hidden neurons.

Finally, we set the desired error $\varepsilon = 10^{-4}$ and the maximum number of steps $MAX_S = 10^4$.

4.3 Analysis of the Experimental Results

Table 1 reports the confusion matrix averaged on the six test sets. The obtained classification accuracy is very good, with an average value of 92%. In some cases the classification accuracy is greater than 95%. As expected, some news and talk shows tend to be confused each other, due to their common cognitive and structural properties. Other false detection results may be due to the high percentage of music with speech in the audio track of commercials, and thus misclassifying some commercials as music clips. In addition, music genre shows the most scattered results, due to its structural, visual and cognitive inhomogeneity.

Table 1. Confusion matrix for the task of TV genre recognition (Unit: 100%)

Genre	*Talk Shows*	*Commercials*	*Music*	*Cartoons*	*Football*	*News*	*Weath.For.*
Talk Shows	**91.7**	0	0	1.6	0	6.7	0
Commercials	1.5	**91**	4.5	0	0	1.5	1.5
Music	1.7	5	**86.7**	5	1.6	0	0
Cartoons	0	0	3.4	**94.9**	0	0	1.7
Football	0	0	0	0	**100**	0	0
News	11.1	0	0	1.6	0	**87.3**	0
Weath.For.	1.5	1.5	0	1.6	0	0	**95.4**

4.4 Comparisons with Other Works

During the recent years many attempts have been done at solving the task of TV genre recognition, according to a number of genres from a reference taxonomy [1,3,4,6,8,12,13,19,20,28,29]. A comparison between our approach and some other classification methods is reported in Table 2. From the analysis of previous works,

several considerations can be made. First of all, the majority of past research focused on either few genres, or well distinguishable genres. The approaches in [1,6,8,20] led to focus on only one kind of genre (either cartoons or commercials). Roach et al [19] presented a method for the classification of videos into three genres (sports, cartoons and news). Dimitrova et al. [3] classified TV programmes according to four genres (commercials, news, sitcoms and soaps). Truong et al. [28] and Xu et al. [29] considered the same set of five genres, which includes cartoons, commercials, music, news and sports. Liu et al. [13] used weather forecasts instead of musics. Dinh et al. [4] split the music genre into two sub-genres (music shows and concerts), thus resulting in a six-genre classifier. As talk shows play an important role in the daily programming of TV channels, we considered the talk show genre in addition to the genres used in [12,13,28,29]. Our classifier can thus distinguish a greater number of genres.

Another common drawback of existing works is that only a restricted set of objects was used as representative of each genre. In fact, typical experimental datasets consisted of few minutes of randomly selected and aggregated audiovisual clips. We believe that this procedure is not applicable in real-world scenarios (e.g. broadcast archives), where users are usually interested in classifying and retrieving objects as whole and complete, rather than as fragments. To overcome this limitation we used complete broadcasted programmes (e.g. an entire talk show), thus limiting potential bias caused by authors in the clip selection phase. A comparison between our experimental dataset and those used in previous works is reported in Table 3. Notice that the total dimension (in minutes) of our experimental dataset increases by a factor of 37 w.r.t. the average dimension of the experimental datasets used in previous works.

A third problem deals with the kind of classifier employed. In many cases, classical statistical pattern recognition methods (i.e. crisp clustering algorithms [4], decision trees [4,28], support vector machines [4,8], Gaussian mixture models [19,29] and hidden Markow models [1,3,13]) 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 (see [11] for details). 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 [7]. Most of the earlier works used the hold-out validation (HOV) technique [3,4,12,13,28,29], 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 [19], 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

Table 2. Classification accuracy compared between our work and some previous works

Authors	*Recognised genres*	*Classifier type*	*Dataset Size [minutes]*	*Data Validation Strategy*	*Classification Accuracy*
Dinh et al. [4]	6	k-NN	110	HOV	96%
Dinh et al. [4]	6	DT	110	HOV	91%
Dinh et al. [4]	6	SVM	110	HOV	90%
Xu et al. [29]	5	GMM	300	HOV	86%
Liu et al. [13]	5	HMM	100	HOV	85%
Truong et al. [28]	5	DT	480	HOV	83%
Liu et al. [12]	5	ANN	100	HOV	71%
Dimitrova et al. [3]	4	HMM	61	HOV	85%
Roach et al. [19]	3	GMM	15	LOOCV	94%
This work	**7**	**MLPs**	**6692**	**KFCV**	**92%**

be highly variable. In our system, we chose to use the K-fold cross-validation (KFCV) of data. This approach limits potential bias that could be introduced in the choice of training and testing data.

Finally, with regard to the accuracy of the experimental results, our approach shows better performance than those in [3,12,13,28,29]. In two cases our approach performs a bit worse [4,19].

Table 3. Details of some experimental databases used for the genre recognition task

Authors	*Clip Duration [seconds]*	*Genre Duration [minutes]*
Dinh et al. [4]	1	news (26), concerts (15), cartoons (19), commercials (18), music shows (11), motor racing games (21).
Xu et al. [29]	300	news (60), commercials (60), sports (60), music (60), cartoons (60).
Liu et al. [13]	1.5	news (20), commercials (20), football (20), basketball (20), weather forecasts (20).
Truong et al. [28]	60	news (96), music (96), sports (96), commercials (96), cartoons (96).
Dimitrova et al. [3]	60	training (26), testing (35).
Roach et al. [19]	30	news (1), cartoons (7), sports (7).
This work	**120 to 2640**	**music (233), commercials (209), cartoons (1126), football (1053), news (1301), talk shows (2650), weather forecasts (120).**

5 Conclusions and Future Work

In this paper we have presented an architecture for the video genre recognition task. Despite the number of existing efforts, the issue of classifying video genres

has not been has not been satisfactory addressed yet. We have overcome the limitations of previous approaches by using *complete* TV programmes (instead of shot clips of content), and by defining a set of features that captures both low-level audiovisual descriptors and higher level semantic information. The greater number of genres considered (w.r.t. previous works), the quality of the experimental dataset built, the type of classifier employed and the data validation technique used makes us conclude that our approach significantly improves the state of the art in the investigated task. Future work will investigate the development of new features, the introduction of new classes (e.g. action movie, comedy, tennis, basketball) and the usefulness of new classifiers.

References

1. Albiol, A., Fullà, M.J.C., Albiol, A., Torres, L.: Commercials detection using HMMs. In: Proc. of the Int. Workshop on Image Analysis for Multimedia Interactive Services (2004)
2. Brugnara, F., Cettolo, M., Federico, M., Giuliani, D.: A system for the segmentation and transcription of Italian radio news. In: Proc. of RIAO, Content-Based Multimedia Information Access (2000)
3. Dimitrova, N., Agnihotri, L., Wei, G.: Video classification based on HMM using text and faces. In: Proc. of the European Conference on Signal Processing (2000)
4. Dinh, P.Q., Dorai, C., Venkatesh, S.: Video genre categorization using audio wavelet coefficients. In: Proc. of the 5th Asian Conference on Computer Vision (2002)
5. Fräba, B., Küblbeck, C.: Orientation Template Matching for Face Localization in Complex Visual Scenes. In: Proc. of the IEEE Int. Conf. on Image Processing, pp. 251–254. IEEE Computer Society Press, Los Alamitos (2000)
6. Glasberg, R., Samour, A., Elazouzi, K., Sikora, T.: Cartoon-Recognition using Video & Audio-Descriptors. In: Proc. of the 13th European Signal Processing Conference (2005)
7. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proc. of the Int. Conf. on Artificial Intelligence (1995)
8. Ianeva, T.I., de Vries, A.P., Rohrig, H.: Detecting cartoons: a case study in automatic video-genre classification. In: Int. Conf. on Multimedia and Expo. (2003)
9. Igel, C., Hüsken, M.: Improving the Rprop Learning Algorithm. In: Proc. of the 2nd Int. ICSC Symposium on Neural Computation (2000)
10. ISO/IEC 15398: Multimedia Content Description Interface (2001)
11. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical Pattern Recognition: A Review. IEEE Trans. on Pattern Analysis and Machine Intelligence 22(1), 4–37
12. Liu, Z., Huang, J., Wang, Y., Chen, T.: Audio feature extraction and analysis for scene classification. In: IEEE Workshop on Multimedia Signal Processing, IEEE Computer Society Press, Los Alamitos (1997)
13. Liu, Z., Huang, J., Wang, Y.: Classification of TV programs based on audio information using Hidden Markov Model. In: Proc. of IEEE Workshop on Multimedia Signal Processing, IEEE Computer Society Press, Los Alamitos (1998)
14. Messina, A., Airola Gnota, D.: Automatic Archive Documentation Based on Content Analysis. IBC2005 Conference Pubblication

15. Messina, A., Montagnuolo, M., Sapino, M.L: Characterizing multimedia objects through multimodal content analysis and fuzzy fingerprints. In: IEEE Int. Conf. on Signal-Image Technology and Internet-Based Systems, IEEE Computer Society Press, Los Alamitos (2006)
16. Montagnuolo, M., Messina, A.: Multimedia Knowledge Representation for Automatic Annotation of Broadcast TV Archives. In: Proc. of the 4th Int. Workshop on Multimedia Semantics (2006)
17. Polikar, R.: Ensemble Based Systems in Decision Making. IEEE Circuits and Systems Magazine 6(3), 21–45 (2006)
18. Quinlan, J.R.: C4.5 - Programs for Machine Learning. Morgan Kaufmann, San Francisco (1993)
19. Roach, M.J., Mason, J.S.D., Pawlewski, M.: Video genre classification using dynamics. In: IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, IEEE Computer Society Press, Los Alamitos (2001)
20. Sánchez, J.M., Binefa, X., Vitriá, J., Radeva, P.: Local Color Analysis for Scene Break Detection Applied to TV Commercials Recognition. In: Huijsmans, D.P., Smeulders, A.W.M. (eds.) VISUAL 1999. LNCS, vol. 1614, pp. 237–244. Springer, Heidelberg (1999)
21. Schwenker, F., Marinai, S.: Artificial Neural Networks in Pattern Recognition. In: Schwenker, F., Marinai, S. (eds.) ANNPR 2006. LNCS (LNAI), vol. 4087, Springer, Heidelberg (2006)
22. Snoek, C.G., Worring, M.: Multimodal video indexing: A review of the state-of-the-art. Multimedia Tools and Applications 25(1), 5–35 (2005)
23. Swain, M.J., Ballard, D.H.: Color indexing. Int. Journal of Computer Vision 7(1), 11–32 (1991)
24. Tamura, H., Mori, S., Yamawaki, T.: Texture features corresponding to visual perception. IEEE Trans. on Systems, Man and Cybernetics 8(6), 460–473 (1978)
25. Taylor, J.S., Cristianini, N.: Support Vector Machines and other kernel-based learning methods. Cambridge University Press, Cambridge (2000)
26. Tekalp, M.: Digital Video Processing. Prentice-Hall, Englewood Cliffs (1995)
27. Tomasi, C.: Estimating Gaussian Mixture Densities with EM – A Tutorial. Duke University (2005)
28. Truong, B.T., Dorai, C., Venkatesh, S.: Automatic Genre Identification for Content-Based Video Categorization. In: Proc. of the 15th IEEE Int. Conf. on Pattern Recognition, IEEE Computer Society Press, Los Alamitos (2000)
29. Xu, L.Q., Li, Y.: Video classification using spatial-temporal features and PCA. In: Proc. of the IEEE Int. Conf. on Multimedia and Expo, IEEE Computer Society Press, Los Alamitos (2003)