Content-based Video Genre Classification Using Multiple Cues

Hazım Kemal Ekenel
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT)
76131 Karlsruhe, Germany
ekenel@kit.edu

Tomas Semela
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT)
76131 Karlsruhe, Germany
semela@student.kit.edu

Rainer Stiefelhagen
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT)
76131 Karlsruhe, Germany
rainer.stiefelhagen@kit.edu

ABSTRACT

This paper presents an automatic video genre classification system, which utilizes several low-level audio-visual cues as well as cognitive and structural information to classify the types of TV programs and YouTube videos. Classification is performed using support vector machines. The system is integrated to our content-based video processing system and shares the same features that we have been using for high-level feature detection task in TRECVID evaluations. The proposed system is extensively evaluated using complete TV programs from Italian RAI TV channel, from French TV channels, and videos from YouTube on which 99.6%, 99%, and 92.4% correct classification rates are attained, respectively. These results show that the developed system can reliably determine TV programs' genre. It also provides a good basis for classifying genres of YouTube videos, which can be improved by using additional information, such as tags and titles, to obtain more robust results. Further experiments indicate that the quality of video does not influence the results significantly. It is found that the performance drop in classifying genres of YouTube videos is mainly due to the large variety of content contained in these videos.

Keywords

Genre classification, audio-visual, TV programs, YouTube videos

1. INTRODUCTION

Automatic video genre classification is an important task in multimedia indexing. Several studies have been conducted on this topic [9]. Initial studies have focused on classifying genres of TV programs. Recently classifying genres of web videos has also attracted significant interest [14, 3, 10, 13]. A challenge "Robust, As-Accurate-As-Human Genre Classification for Video" has also been included within Multimedia Grand Challenge [1].

AIEMPro'10 Florence, Italy

The studies on TV genre classification mainly utilize audio-visual features. For example, in [7] color statistics, cut detection, camera motion, object motion, and audio are extracted from videos. From these properties further style attributes like camera panning and zooming, scene transitions, object motion, speech, and music are extracted. Finally the video is classified as news, car race, tennis, commercials, or cartoons, using its style profile, derived from the style attributes. In [12], they use camera motion and a dominant color feature as visual features and mel-frequency cepstral coefficients as audio features. The classification is performed using pseudo 2D hidden Markov models.

Besides audio-visual information, the studies on web video genre classification also benefit from additional information such as tags, titles, and surrounding text. In [14], semantic features based on concept histograms and text features that are derived from the title, tags, and video description are used. Similarly, in the Multimedia Grand Challenge on genre classification [1], semantic features, titles, tags, relevance from related videos, and user interest deduced from the user's videos are exploited [3, 10, 13].

A comprehensive overview of the studies on TV genre classification can be found in [8, 9]. In these papers, the authors also present a robust TV genre classification system. In their first system [8], they utilize visual features based on color —hue, saturation, value, luminance—, texture —Tamura features contrast, directionality—, and temporal activity. The audio features consist of average speech rate and normalized duration values of several audio classes, e.g. silence, noise, and speech. They also utilize *structural features*, that contain average shot length and distribution of the shot lengths along the video, and *cognitive features* that contain information about average number of faces per video and their position. They use a multi-layer perceptron for classification. In this study, the authors built a very large dataset containing 262 complete TV programs of around 110 hours. They achieved 92% accuracy over seven genres. The work in [9] extends their previous work with modifications in *structural* and *cognitive* features. With these modifications, their performance increased from 92% to 95% on the same dataset.

In this study, we developed an automatic video genre classification system which uses the features that are already extracted for detecting high-level features (HLF) in videos [5, 6]. This brings two main advantages. The first advantage is since the features are already extracted for HLF detection, there is only a classification step as an overhead to our

content analysis system. The other advantage is that knowing the genre of the processed video could give cues about the available high-level features in the video, which in turn could lead to a higher performance in the high-level feature detection task. In our study, the used low level visual features are: HSV color histogram, color moments, autocorrelogram, co-occurrence texture, wavelet texture grid, and edge histogram. Mel-frequency cepstral coefficients, fundamental frequency, signal energy, and zero crossing rate are the features extracted from audio. In addition to these audio-visual features, cognitive and structural features as suggested in [8] are also exploited. Extensive experiments have been conducted on three different datasets. The experimental results on Italian and French TV programs have shown that the system can reliable classify genres of TV programs. It has been observed that classifying genres of YouTube videos poses a more difficult problem due to the diversity in the content of these videos. However, using audio-visual features still provides a robust basis, whereas further improvement in the performance could be realized by using a training dataset that covers more variety and exploiting additional information, such as tags.

The organization of the paper is as follows. In Sections 2 and 3, low level visual features and audio features are briefly explained, respectively. Cognitive and structural features are shortly described in Section 4. In Section 5, information about the classification process is conveyed. Experimental results are presented and discussed in Section 6. Finally, in Section 7, conclusions are given.

2. LOW LEVEL VISUAL FEATURES

We used six different low level visual features, which represent color and texture information in the video. These are the features that we also utilize for content analysis to detect high-level features in the videos.

2.1 Color descriptors

Color features are one of the most popular visual features in the area of image retrieval, since color features are less dependent on the size, direction, and view point of images compared to other visual features. We use three different types of color descriptors.

HSV histogram.

We opt for the HSV color space and build a histogram with 162 bins. We quantized the Hue (H) values, which represent the color information, more precisely. We assigned 18 bins for the "H" channel, 3 bins for the saturation (S) channel, and 3 bins for the value (V) channel ($18 \times 3 \times 3 = 162$).

Color moments.

The first three color moments have been used. We divide an image into $k \times k$ blocks, and extract color moments from each image block. The final feature vector is obtained by concatenating the color moments extracted from the blocks, which results in a $9 \times k \times k$ feature vector. Here we set $k = 5$ resulting in a 225-dimensional feature vector.

Autocorrelogram.

The color correlogram was proposed to characterize not only the color distributions of pixels, but also the spatial correlation between pairs of colors.

If we consider all the possible combinations of color pairs the size of the color correlogram will be very large. Therefore a simplified version of the feature called the color autocorrelogram is often used instead. The color autocorrelogram only captures the spatial correlation between identical colors and thus reduces the dimension to $O(Nd)$.

64 quantized color bins and five distances are used for this representation.

2.2 Texture descriptors

Texture features are also an important group of image descriptors. We use three different types of texture descriptors.

Co-occurrence texture.

Five types of features extracted from the gray level co-occurrence matrix (GLCM): Entropy, Energy, Contrast, Correlation, and Local homogeneity. Those features are extracted from 24 different GLCMs, in our case with 8 gray level bins, at different orientations and distances. The resulting vector is $24 \times 5 = 120$-dimensional.

Wavelet texture grid.

The implementation follows the description in [4], obtaining the variances of the high-frequency sub-bands of the wavelet transform of each grid region. We used 12 sub-bands (4-level analysis). The used wavelet basis function is the simple Haar wavelet while the grid has $4 \times 4 = 16$ regions. Thus, the resulting vector is $16 \times 12 = 192$-dimensional.

Edge histogram.

For the edge histogram, 5 filters as proposed in the MPEG-7 standard are used to extract the kind of edge in each region of 2×2 pixels. Then, those small regions are grouped in a certain number of areas (4 rows $\times$ 4 columns in our case) and the number of edges matched by each filter (vertical, horizontal, diagonal 45°, diagonal 135° and non-directional) are counted in the region's histogram. Thus, the resulting vector is $4 \times 4 \times 5 = 80$-dimensional.

3. AUDIO FEATURES

Audio plays a crucial role in human perception to recognize or classify different programs. Therefore it is of paramount importance to utilize audio information in the genre classification task.

At the moment four features are computed from the audio signal of every video for classification. Additional features can be integrated easily in the future.

All features are computed from a mono-channel, uncompressed PCM audio signal with a 16kHz sample rate and a 256 kbit/s bit rate.

The single features are computed over small non-overlapping windows of *N=320* samples using the Hamming window function. In the following equations m is the index of the window and $s_a(n)$ is the signal at the time index n.

Mel-frequency cepstral coefficients.

MFCCs are commonly used features in music similarity tasks and speech recognition. The 8^{th} order mel-frequency cepstral coefficients are computed in this study.

Fundamental Frequency.

The fundamental frequency is defined as the inverse of the

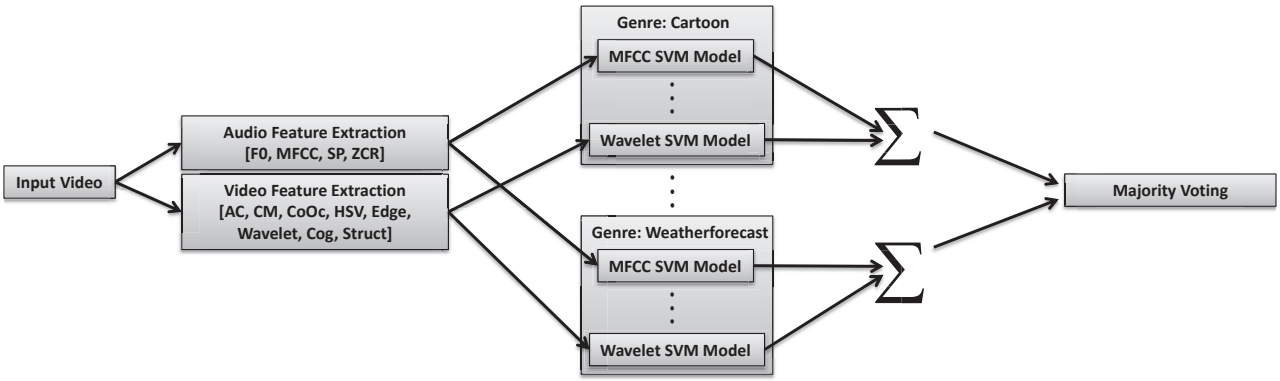

Figure 1: System Overview

frame length of a periodic or quasi periodic signal:

$$F_0(m) = \frac{1}{T_{min}(m)}$$

Signal Energy.

Signal energy is defined as the mean square of the amplitude in the current window:

$$SP(m) = \sum_{n=m-N+1}^{m} s_a(n)^2$$

Zero Crossing Rate.

The zero crossing rate measures the rate of zero crossings in the amplitude of the signal, averaged by the length of the frame length:

$$ZCR(m) = \frac{1}{N} \sum_{n=m-N+1}^{m} \frac{|sign(s_a(n) - sign(s_a(n-1))|}{2}$$

Two possible feature vector representations were chosen for the experiments:

1. A single 22-dimensional feature vector consisting of the mean and standard deviation for each feature over the whole audio signal.

2. Each feature is saved into a separate file. Mean and standard deviation of each feature are computed over a time frame of 1 second for the whole audio signal. These values are used as an input to compute a 10-component Gaussian mixture model resulting in a feature vector containing mean, standard deviation and weight for each Gaussian function. Since the input data for *ZCR*, *SP* and *F0* are 2-dimensional and in case of MFCCs 16-dimensional the four feature vectors are 50- and 330-dimensional, respectively.

4. COGNITIVE AND STRUCTURAL FEATURES

Cognitive and structural features are implemented as proposed in [8]. Cognitive feature is derived using a face detector [11]. It contains average number of faces per frame, distribution of number of faces per frame and distribution of location of the faces in the frame.

Structural feature is derived using a shot boundary detector [5]. It contains average shot duration and distribution of shot lengths.

Table 1: Number of genre videos in the data sets

	Italian TV	French TV	YouTube
Cartoon	27	-	60
Commercial	58	-	60
Football	22	3	10
Music Show	7	18	60
News	49	30	60
Talk Show	39	27	60
Weather For.	60	-	60
Total	262	78	370

5. CLASSIFICATION

Classification is performed using support vector machine (SVM) classifiers. Radial basis function is used as the kernel. One-vs-all strategy is employed to train a SVM for each feature and each genre.

The training is conducted using a cross-fold validation scheme to determine the optimal penalty parameter C and the radial basis function parameter γ for each SVM. Various combinations of the (C, γ) pairs are tried and the one with the best cross-validation accuracy is picked. The best combination is then used to train the whole training set.

During testing, the audio-visual features are extracted from the input video and each feature is classified using the corresponding SVM model from each genre. The outputs of each SVM model in a genre class are then combined using sum rule. The final classification decision is then taken via majority voting, that is by selecting the genre class with the highest score. Overview of the system is depicted in Figure 1.

6. EXPERIMENTS

We used three different datasets to evaluate the proposed automatic video genre classification system. Experiments have been conducted both on complete TV programs and videos from YouTube. Number of videos and represented genres can be viewed in Table 1.

Table 2: Average correct classification rates obtained on the RAI dataset

System	Class. Rate
RAI07 [8]	92.0%
RAI09 [9]	94.9%
KIT	99.6%

Table 3: Confusion matrix obtained on the RAI dataset (%)

	Ca	Co	Fo	Mu	Ne	Ta	We
Ca	100	0	0	0	0	0	0
Co	0	100	0	0	0	0	0
Fo	0	0	100	0	0	0	0
Mu	0	14.2	0	85.7	0	0	0
Ne	0	0	0	0	100	0	0
Ta	0	0	0	0	0	100	0
We	0	0	0	0	0	0	100

6.1 Experiments on the RAI dataset

RAI dataset is the largest available TV genre dataset, which contains TV programs from Italian TV channel RAI [8]. It also enables us to compare our results with the state-of-the-art TV genre classification systems [8, 9]. The dataset contains 262 complete TV programs of around 110 hours belonging to seven genres —cartoons, commercials, football, music, news, talk shows, and weather forecasts—. As in [8, 9], we divided the dataset into six subsets. We run six experiments by using each time five of the subsets for training and one subset for testing.

Average correct classification rates over these six experiments are presented in Table 2. In the table, RAI07 corresponds to the system in [8], RAI09 corresponds to the system in [9], and KIT corresponds to the proposed system. Our system outperforms both of the state-of-the-art systems. The only error was due to confusion of music program with a commercial as can be seen from the confusion matrix in Table 3. Due to limited space, only first two letters of each genre are used in this table. Classification accuracy of each four components separately and overall accuracy can be viewed in Figure 2. The reason of superior performance is mainly due to the used low-level visual features' capability to discriminate the TV genres. As can be seen from Figure 3, individual color and texture features are able to reach over 90% classification rate. HSV histogram, color moments, edge histogram, and wavelet features are found to be the most discriminative features reaching and exceeding 95% classification rate. The cognitive and structural features adopted from [8] provide the least contribution to the overall performance.

Figure 4 shows the individual audio feature results using the second feature representation with GMMs. The mel-frequency cepstral coefficients show the most promising results. Overall the audio feature accuracy was 86.6% with the GMM representation of audio features, whereas the first representation of the audio feature vector achieved an accuracy of 96%. The combination with the visual features lead to an overall accuracy of 99.2% and 99.6%, in the case of first and second audio feature representation, respectively. The four less accurate audio feature vectors contribute slightly more to the overall accuracy than the single audio feature vector with better accuracy.

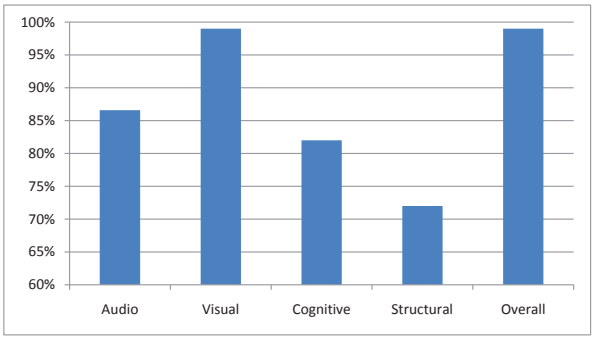

Figure 2: Performance of the four main components on the RAI dataset. Audio features are modeled with GMM.

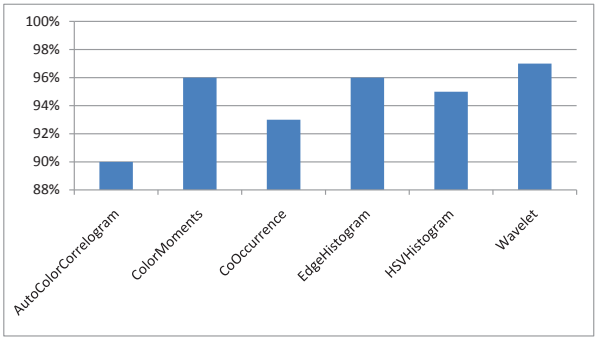

Figure 3: Performance of visual components on the RAI dataset

6.2 Experiments on the Quaero dataset

We also tested the proposed approach on the dataset from Quaero project [2]. The dataset contains 78 TV programs from French TV channels that belong to four different genres: football, music, news, and talk show. Quaero evaluation guidelines are followed and 3-fold testing is performed. In the experiments 99% average correct classification rate is achieved. Only in one of the folds a music program was classified as a talk show program. This result validates that the proposed approach is able to classify genres of TV programs reliably. Individual accuracies of the audio features are 93% for the single audio feature and 82% for the GMM modeled seperate audio features. The visual cues alone reached an accuracy of 99%. Adding either of the audio representation did not improve the results any further.

6.3 Experiments on the YouTube dataset

We collected 370 YouTube videos of around 40 hours belonging to the same seven genres as the ones in the RAI dataset. To collect the dataset, we first used the genre names as search queries. Afterwards the downloaded videos were manually checked whether they really belong to the searched genre or not. As in the experiments on the RAI dataset, 6-fold testing is performed. As can be seen from the first row of Table 4, 92.4% average classification rate is attained in these experiments. Lower performance indicates difficulty of doing genre classification on YouTube videos. In order

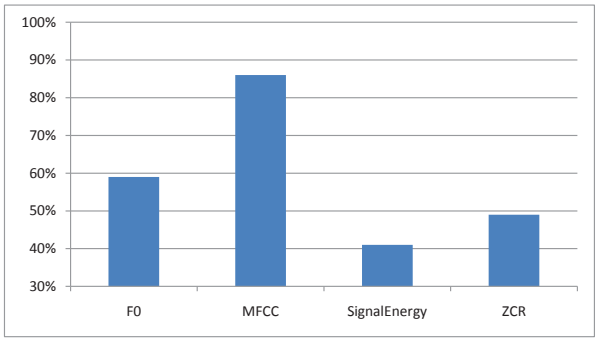

Figure 4: Performance of audio components on the RAI dataset. Audio features are modeled with GMM.

to assess whether the main cause of performance loss is due to high variety of the videos or the low video quality, we encoded the TV programs in the RAI dataset to low quality YouTube format and performed genre classification experiments. As can be noticed from the second row of Table 4, the proposed system achieves 99.2% on the RAI dataset encoded to a low quality. This result shows that the system is able to classify the genres of low quality videos and the main problem is high variety contained in YouTube videos. To account for the high variety among the programs that belong to the same genre in YouTube, more training data can be used. Moreover, by using additional available information, such as tags and surrounding text, the performance can be further improved.

Figure 5 shows the performance of the individual components and the combined system on the YouTube videos. Without using audio information, 81.6% correct classification rate is achieved. Addition of audio information improves the performance to 92.4%, which uses the second audio feature representation with an individual accuracy of 70.5%. The first audio feature representation leads to 70% correct classification rate. Its combination with visual features increases the performance to 89.2%. The audio classification proves to be a significant contribution when classifing YouTube videos. Note that, while classifying the genres of TV programs, the addition of audio features did not provide a significant improvement in performance since the correct classification rate obtained by visual features only were already very high. However, in the case of genre classification of YouTube videos, which is a more difficult task and at which using only visual features cannot lead to high classification accuracy, the contribution of audio features to the performance is more visible. Also the mel-frequency cepstral coefficients again obtained the highest accuracy among the audio features. Another interesting observation from Figures 5, 6, and 7 is the relative performance of the components and the features of the components stay the same over different datasets.

The confusion matrix is given in Table 5. Due to space limit only first two letters of each genre are used in the table. The most difficult genres to classify correctly are found to be music and football, which are often confused with cartoons and commercials. Cartoons are found to be the easiest genre to discriminate with 100% accuracy.

Table 4: Average correct classification rates obtained on YouTube videos and RAI videos encoded in low quality

Training	Testing	Class. Rate
YouTube	YouTube	92.4%
RAI LQ	RAI LQ	99.2%

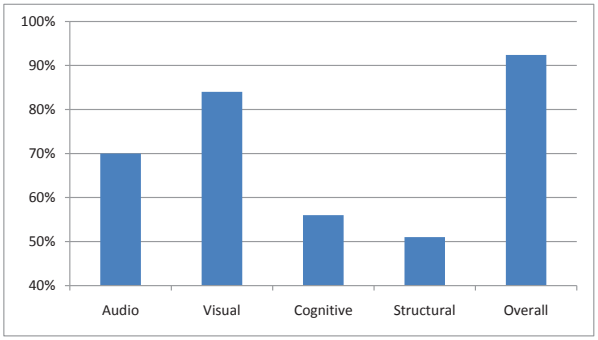

Figure 5: Performance of the four main components on the YouTube dataset. Audio features are modeled with GMM.

7. CONCLUSIONS

In this paper we presented a content-based automatic video genre classification system. The proposed system exploits low level audio and visual features and combines them with cognitive and structural information. In the system, as low level visual features, color and texture descriptors are used. Cognitive information is based on the number of faces per frame and their distribution over video and different locations. Structural information contains average shot duration and distribution of shot lengths. Mel-frequency cepstral coefficients, fundamental frequency, signal energy, and zero crossing rate are computed to represent the audio information of each video.

The proposed approach has been extensively tested on two different TV program datasets and on YouTube videos. Very high correct classification rates are achieved on TV programs: 99.6% on the RAI dataset and 99% on the Quaero dataset. It has been found that due to high variation in content, it is more difficult to classify genres of YouTube videos. The achieved correct classification rate on these videos is 92.4%. For example, some YouTube football videos show a best-of some football player with music in the background, which makes it more difficult to distinguish it from a commercial or music video. However, with the use of training data that can cover more variety and utilizing text information, the performance is expected to match the ones obtained on the TV program datasets.

Examining each feature individually, color moments, wavelet texture grid, and edge histogram are found to be most useful visual cues; whereas mel-frequency cepstral coefficients have proven to be the most promising audio feature.

In summary, the developed system performs genre classification very efficiently without causing a significant computational load to our high-level feature detection system. Moreover, obtained genre information is expected to provide additional cues which can be used to improve the HLF detection system's performance.

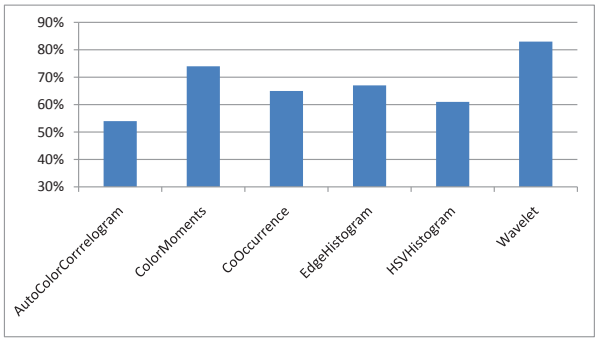

Figure 6: Performance of visual components on the YouTube dataset

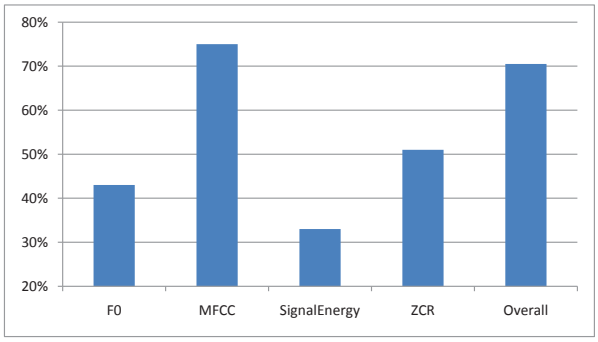

Figure 7: Performance of audio components on the YouTube dataset. Audio features are modeled with GMM.

Acknowledgments

The authors would like to thank Alberto Messina and Maurizio Montagnuolo from RAI Centre for Research and Technological Innovation for their contributions to the study and for providing the TV program data. The authors also thank Mika Fischer and Hua Gao for their contributions to the study. This study is funded by OSEO, French State agency for innovation, as part of the Quaero Programme.

8. REFERENCES

[1] Multimedia Grand Challenge website *http://comminfo.rutgers.edu/conferences/mmchallenge/*.

[2] Quaero Programme website *http://www.quaero.org/*.

[3] D. Borth, et al., "TubeFiler – an Automatic Web Video Categorizer." *Proc. of ACM Multimedia*, pp. 1111–1112, Beijing, China, 2009.

[4] M. Campbell, et al., "IBM Research TRECVID-2006 Video Retrieval System", *NIST TRECVID Workshop*, Gaithersburg, USA, 2006.

[5] H.K. Ekenel, M. Fischer, H. Gao, K. Kilgour, J.S. Marcos, R. Stiefelhagen, "Universität Karlsruhe (TH) at TRECVID 2007", *NIST TRECVID Workshop*, Gaithersburg, MD, USA, 2007.

[6] H.K. Ekenel, H. Gao, R. Stiefelhagen, "Universität Karlsruhe (TH) at TRECVID 2008", *NIST TRECVID Workshop*, Gaithersburg, MD, USA, 2008.

[7] S. Fischer, R. Lienhart, W. Effelsberg, "Automatic Recognition of Film Genres", *Proc. of ACM Multimedia*, pp. 295–304, San Francisco, USA, 1995.

[8] M. Montagnuolo, A. Messina, "TV Genre Classification Using Multimodal Information and Multilayer Perceptrons" , *AIIA, LNAI 4733*, pp. 730-741, 2007.

[9] M. Montagnuolo, A. Messina, "Parallel Neural Networks for Multimodal Video Genre Classification" , *Multimedia Tools and Appl.*, 41(1):125–159, 2009.

[10] Y. Song, Y. Zhang, X. Yhang, J. Cao, J. Li, "Google Challenge: Incremental-learning for Web Video Categorization on Robust Semantic Feature Space", *Proc. of ACM Multimedia*, pp. 1113–1114, Beijing, China, 2009.

[11] P. Viola, M. J. Jones, "Robust real-time face detection", *Intl. Journal of Computer Vision*, 57(2):137–154, 2004.

[12] J. Wang, C. Xu, E. Chng, "Automatic Sports Video Genre Classification Using Pseudo-2D-HMM", *Proc. of Intl. Conf. on Pattern Recognition*, pp. 778–781, Washington DC, USA, 2006

[13] X. Wu, W.L. Zhao, C.W. Ngo, "Towards Google Challenge: Combining Contextual and Social Information for Web Video Categorization", *Proc. of ACM Multimedia*, pp. 1109–1110, Beijing, China, 2009.

[14] L. Yang, J. Liu, X. Yang, X.S. Hua, "Multi-Modality Web Video Categorization", *Proc. of Multimedia Information Retrieval, MIR '07*, pp. 265–274, Augsburg, Germany, 2007.

Table 5: Confusion matrix obtained on the YouTube dataset (%)

	Ca	Co	Fo	Mu	Ne	Ta	We
Ca	100	0	0	0	0	0	0
Co	0	95.0	3.3	1.6	0	0	0
Fo	0	10	80	10	0	0	0
Mu	5	8.3	0	85.0	0	1.6	0
Ne	0	1.6	0	0	90.0	6.6	1.6
Ta	0	0	0	5	1.6	93.3	0
We	3.3	0	0	0	0	3.3	93.3