CARTOON-RECOGNITION USING VIDEO & AUDIO DESCRIPTORS

Ronald Glasberg, Amjad Samour, Khalid Elazouzi and Thomas Sikora

Communication Systems Group, Technical University of Berlin
Einsteinufer 17, 10587 Berlin, Germany
phone: + (49) 30-314 28932, fax: + (49) 30-314 22514, email: [glasberg, samour, sikora]@nue.tu-berlin.de
web: http://www.nue.tu-berlin.de/

ABSTRACT

We present a new approach for classifying mpeg-2 video sequences as ‘cartoon’ or ‘non-cartoon’ by analyzing specific video and audio features of consecutive frames in real-time. This is part of the well-known video-genre-classification problem, where popular TV-broadcast genres like cartoon, commercial, music, news and sports are studied. Such applications have also been discussed in the context of MPEG-7 [12]. In our method the extracted features from the visual descriptors are non-linearly combined using a multilayered perceptron and then considered together with the output of the audio-descriptor to produce a reliable recognition. The results demonstrate a high identification rate based on a large collection of 100 representative video sequences (20 cartoons and 4*20 non-cartoons) gathered from free digital TV-broadcasting.

1. INTRODUCTION

With the advent of digital TV-broadcasts presenting more than hundred of channels at a time, the need for a user-friendly TV-program selection is growing. Unlike the present TV, a new system should enable users to access programs clustered by genres. The main goal of our research is therefore the classification of an mpeg-2 video-stream in real-time into genres at the highest level.

2. RELATED WORK

The recent approaches addressing cartoon-detection and video-classification are listed in [1-3] and [4-9] respectively. These methods extract a number of different low-level features, from which the analysis is made to build so-called signatures to describe a certain video class.
Roach et al. published an approach for the classification of sequences as cartoons using only one descriptor [2] extracting a motion-feature on a database of 8 cartoons and 20 non-cartoon sequences all together of 20 minutes and later on extended his method to the genre-classification [5] problem. Athitsos et al. [3] emphasized the basic characteristics of cartoons and implemented nine colour descriptors in order to distinguish photographs and graphics on the www on a database of 1200 samples. In the same fashion, but with more insight into the basic characteristics of cartoons, Ianeva et al. [1] implemented six descriptors. It is difficult to predict how the described systems would perform on a test set of mpeg-2 streams.
Concerning audio, a variety of features has been proposed in the literature for characterizing signals [13]. Generally they can be divided into two categories: physical features and perceptual features. The perceptual feature describes the perception of sounds by human beings. Loudness, pitch and timbre are examples of these features. The physical features such as zero crossing-rate, MFCC and energy are further grouped into spectral features and temporal features according to the domain in which they are calculated [14].
To recognize a genre, only a few of these features are useful. The Mel-scale Frequency Cepstral Coefficients MFCC approach is widely used in the area of speech recognition; it is also one of the most used features for audio classification.

We use a database of 100 representative video sequences (e.g. cartoons with dark frames, commercials with animated cartoon sequences) and five new/ modified visual descriptors together with a selected audio descriptor.

3. GENRE-CLASSIFICATION PROCESS

In a first step each mpeg-2 stream is divided into the visual and acoustical data.

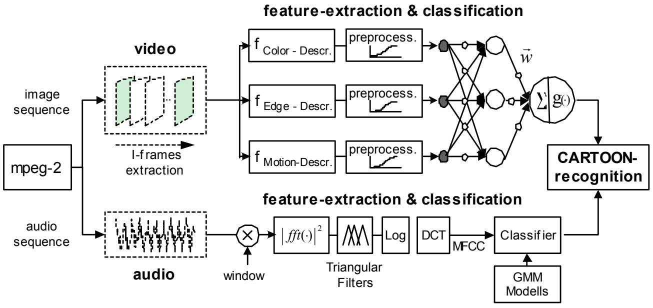


Figure 1: Process involved computing the feature vectors

We consider consecutive I-frames of the image sequence and the corresponding audio frames.

3.1 Video-Descriptors

By looking at frames from cartoon and non-cartoon sequences we observed basic differences between them, like the appearance of bright, highly saturated colors, areas of uniform color, few sharp edges and a low motion activity. Based on these observations, we designed five descriptors to transform our data into feature vectors.

3.1.1. Brightness-Descriptor family

We take each frame and determine the average brightness as well as the amount of pixels with a level higher than a threshold Th_L. These values are averaged over a window N of frames resulting in f_1 and f_2 respectively.

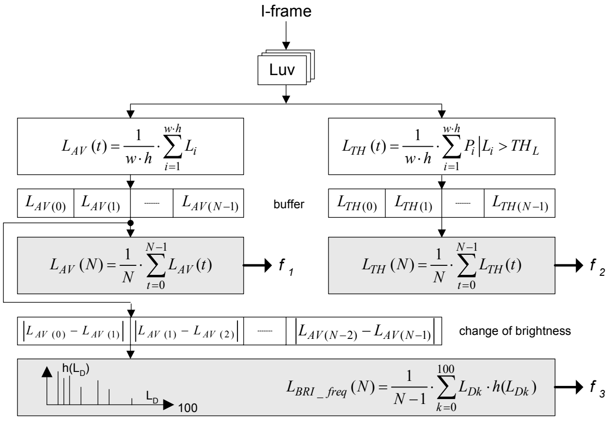

I-frame
Luv
$L_{AV}(t)=\frac{1}{w\cdot h}\cdot\sum_{i=1}^{w\cdot h}L_i$
$L_{TH}(t)=\frac{1}{w\cdot h}\cdot\sum_{i=1}^{w\cdot h}P_i\left|L_i>TH_L\right.$
$L_{AV(0)}$ $L_{AV(1)}$ — $L_{AV(N-1)}$
buffer
$L_{TH(0)}$ $L_{TH(1)}$ — $L_{TH(N-1)}$
$L_{AV}(N)=\frac{1}{N}\cdot\sum_{t=0}^{N-1}L_{AV}(t)$
f_1
$L_{TH}(N)=\frac{1}{N}\cdot\sum_{t=0}^{N-1}L_{TH}(t)$
f_2
$|L_{AV(0)}-L_{AV(1)}|$ $|L_{AV(1)}-L_{AV(2)}|$ — $|L_{AV(N-2)}-L_{AV(N-1)}|$
change of brightness
$h(L_D)$
L_D
100
$L_{BRI\_freq}(N)=\frac{1}{N-1}\cdot\sum_{k=0}^{100}L_{Dk}\cdot h(L_{Dk})$
f_3

Figure 2: Block diagram of the brightness-descriptor

Our interest is also focused on the change of brightness for consecutive frames; therefore we developed $L_{Bri\text{-}freq}$ with the output f_3.

3.1.2. Saturation-Descriptor family

The saturation-descriptor uses the HSV-Space. Similar to the brightness-descriptor we determine the change of saturation f_5 for consecutive frames.

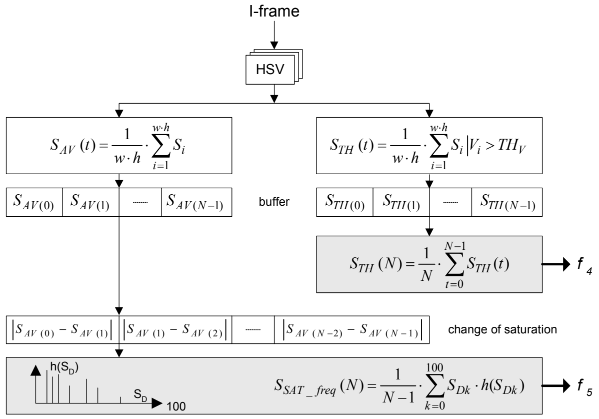

I-frame
HSV
$S_{AV}(t)=\frac{1}{w\cdot h}\cdot\sum_{i=1}^{w\cdot h}S_i$
$S_{TH}(t)=\frac{1}{w\cdot h}\cdot\sum_{i=1}^{w\cdot h}S_i\left|V_i>TH_V\right.$
$S_{AV(0)}$ $S_{AV(1)}$ — $S_{AV(N-1)}$
buffer
$S_{TH(0)}$ $S_{TH(1)}$ — $S_{TH(N-1)}$
$S_{TH}(N)=\frac{1}{N}\cdot\sum_{t=0}^{N-1}S_{TH}(t)$
f_4
$|S_{AV(0)}-S_{AV(1)}|$ $|S_{AV(1)}-S_{AV(2)}|$ — $|S_{AV(N-2)}-S_{AV(N-1)}|$
change of saturation
$h(S_D)$
S_D
100
$S_{SAT\_freq}(N)=\frac{1}{N-1}\cdot\sum_{k=0}^{100}S_{Dk}\cdot h(S_{Dk})$
f_5

Figure 3: Block diagram of the saturation-descriptor

For the threshold f_4 we consider only the saturation with a brightness-value higher than a threshold Th_V.

3.1.3. Color Nuance Descriptor

We determine the mean color distance of each pixel with his adjacent eight neighbours (except border); calculate the mean value for the frame, over window N of frames and obtain f_6.

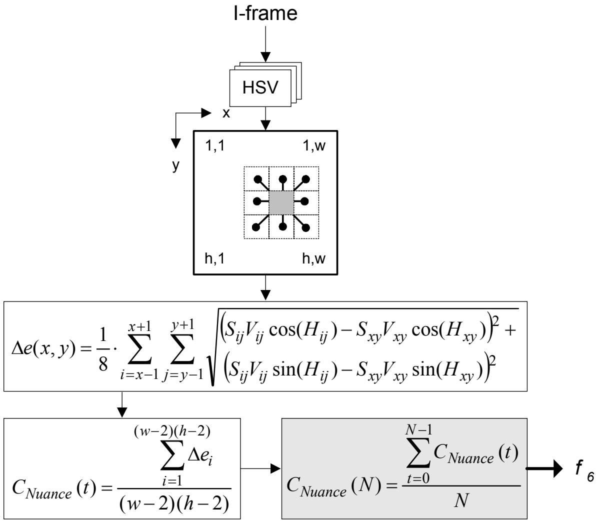

I-frame
HSV
x
y
1,1
1,w
h,1
h,w
$\Delta e(x,y)=\frac{1}{8}\cdot\sum_{i=x-1}^{x+1}\sum_{j=y-1}^{y+1}\sqrt{\left(S_{ij}V_{ij}\cos(H_{ij})-S_{xy}V_{xy}\cos(H_{xy})\right)^2+\left(S_{ij}V_{ij}\sin(H_{ij})-S_{xy}V_{xy}\sin(H_{xy})\right)^2}$
$C_{Nuance}(t)=\frac{\sum_{i=1}^{(w-2)(h-2)}\Delta e_i}{(w-2)(h-2)}$
$C_{Nuance}(N)=\frac{\sum_{t=0}^{N-1}C_{Nuance}(t)}{N}$
f_6

Figure 4: Block diagram of the color-nuance-descriptor

3.1.4. Edge-Descriptor

We detected, that Cartoons have in comparison to the other mentioned genres in general less sharp edges. Therefore we implemented the Canny detector [10]. In our experimental tool the result is labeled with f_7.

3.1.5. Motion-Descriptor family

We implemented one descriptor [11] which extracts the motion-activity information included in the mpeg-2 stream as well as one descriptor extracting the motion information f_8 of a video-stream similar to [2].

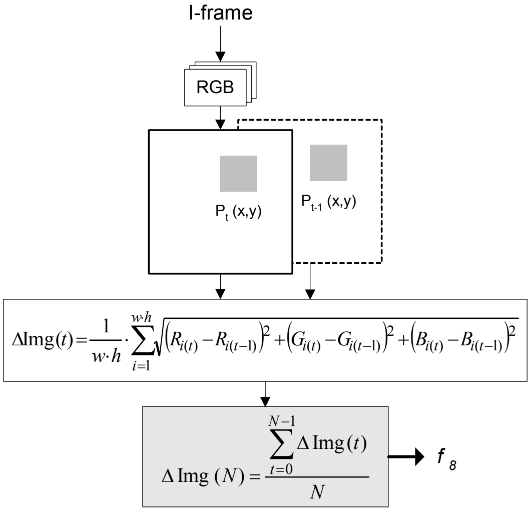

I-frame
RGB
$P_t(x,y)$
$P_{t-1}(x,y)$
$\Delta\mathrm{Img}(t)=\frac{1}{w\cdot h}\cdot\sum_{i=1}^{w\cdot h}\sqrt{\left(R_{i(t)}-R_{i(t-1)}\right)^2+\left(G_{i(t)}-G_{i(t-1)}\right)^2+\left(B_{i(t)}-B_{i(t-1)}\right)^2}$
$\Delta\,\mathrm{Img}(N)=\frac{\sum_{t=0}^{N-1}\Delta\,\mathrm{Img}(t)}{N}$
f_8

Figure 5: Block diagram of the motion-descriptor

3.2 Audio-Descriptor

The extraction method of MFCC is presented in the lower part of Figure 1. The audio signal is divided into windowed frames. After a Fast-Fourier-Transformation FFT, the power spectrum is transformed to mel-frequency scale using a filter bank consisting of triangular filters. Finally, the discrete cosine transform DCT of the logarithm is performed to calculate the cepstral coefficients from the mel-spectrum. The MFCC are given by:

$$c_i = \sum_{k=1}^{K} \log(S_k) \cdot \cos\left(\frac{i\pi}{K}\left(k - \frac{1}{2}\right)\right) \quad \text{for } i = 1,2,......K\,,$$

where c_i is the i^{th} MFCC, S_k is the output of the k^{th} filter bank channel and K is the number of coefficients.

4. EXPERIMENTS

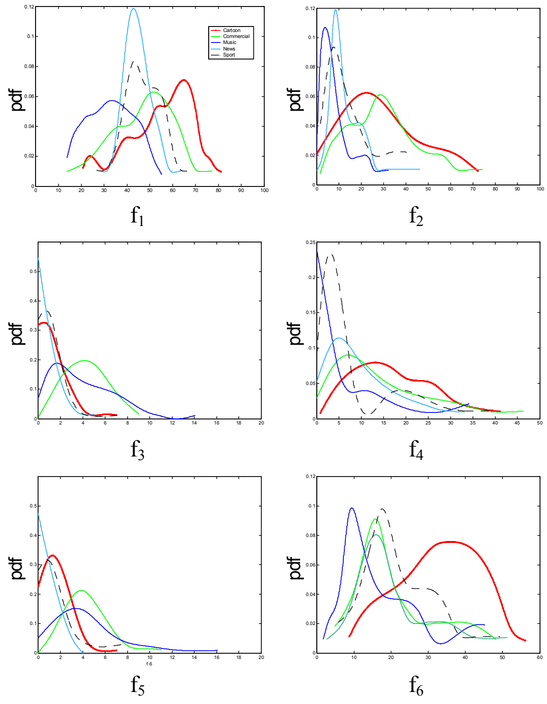
The experiments were carried out on a representative large collection of mpeg-2 video sequences in total of 200 min of recordings; 20 cartoons and 4*20 non-cartoons (commercial, music, news and sport) of 2 minutes' each gathered from popular TV broadcasting. We used the following parameters:

- Video-Feature-Extraction: We extracted consecutive frames and scaled them down to a resolution of 90*72 pixels. The number of frames in the pre-processing window is N=50.

- Audio-Feature-Extraction: The audio signals from the corresponding videos are divided into sub-segments of 1s length without overlapping. To apply the feature extraction, each sub-segment is divided into 40ms frames with a 50 % overlap between adjacent frames, then multiplied by the hamming-window function and transformed into frequency domain using FFT. We extracted 13 static MFCC-features and their first and second derivatives, so that our feature vector has finally a dimension of 39.

4.1 Experimental Results

Figure 6 shows the 8 pdf's probability density functions for the visual descriptor f_1 … f_8.



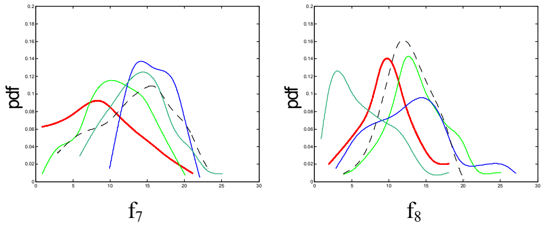


Figure 6: Results of the descriptors f_1-f_8 for each genre

The recognition of cartoons with the extracted feature of a single video-descriptor is obviously not sufficient. Therefore we combined the descriptors with a multilayered perceptron. The results of windows N=50 were averaged over the whole length of each video sequence. Fig. 7 depicts the detection rates for the 20 video sequences of each genre used in our experiments.

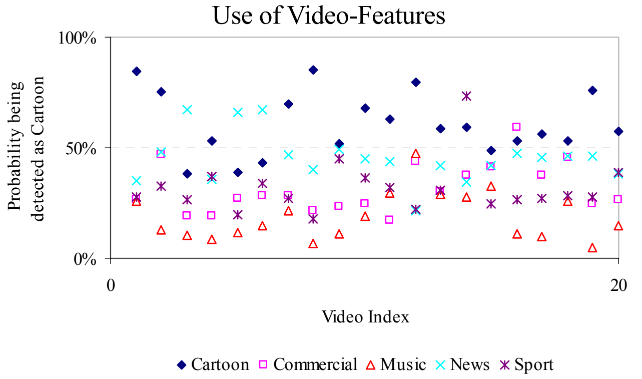


Figure 7: Probability for videos being detected as Cartoon

In our experiment 16 videos from the cartoon-genre were classified as cartoons. The remaining 4 videos were very close to the cartoon-detection threshold main even they included long dark shots and many non-uniform colour areas. It is obvious, that detection of any genre from dark image sequences is highly unreliable. An increase of detection is possible by taking the audio information into account Fig. 8.

The recognition using the audio-mode is possible every second. In order to synchronize the audio with the video part, the result of each second were averaged over a length of 25s.

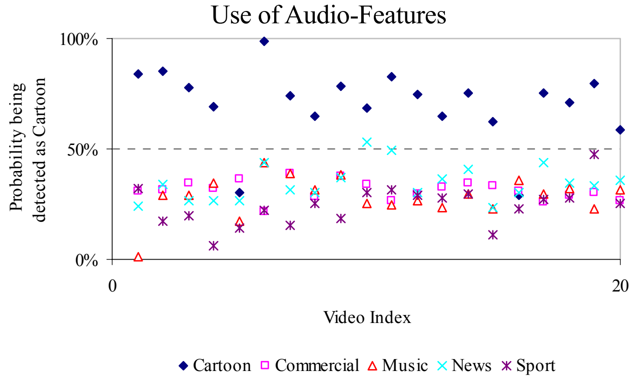


Figure 8: Probability for videos being detected as Cartoon using the extracted audio-features

It is interesting to note from Fig. 7, that with the developed Video-Descriptors an average correct classification rate of 80% for cartoon-videos detected as a ‘cartoon’ and more than 85% for news, 95% for commercial, sport and nearly 100% of music-videos detected as ‘non-cartoon’ has been achieved. From Fig. 8 it can be seen, that using the Audio-Descriptor (training set 70% and testing set 30%) an even higher detection rate, namely 90% for cartoon-videos detected as a ‘cartoon’ has been received.

In a first step, we used a threshold for the results of the Video and Audio-Descriptors respectively and combined them linearly. Table 1 shows a summary for the experiment results.

	Recognition as Cartoon		
Genre	Visual-Descriptor	Audio-Descriptor	Video & Audio
Cartoon	80%	90%	90%
Commercial	5%	0%	0%
Music	0%	0%	0%
News	15%	5%	0%
Sports	5%	0%	0%

Table 1: Classification accuracy on the database of 100 mpeg-2 streams

This result demonstrates that the performance of the Video and Audio-Descriptors is improved through a suitable combined audio-video classification.

5. SUMMARY & CONCLUSION

We have presented in this paper a nonlinear approach for the detection of cartoons. Three key contributions have been made. We started with the development of new or modified visual descriptors. Second, we developed and applied a nonlinear MLP to combine the results of the visual descriptors. Third we considered the audio-features and combined them linearly with the concerning video-features.

A video database containing five popular genres namely cartoon, commercial, music, news and sports has been used. An average correct classification rate of 90% for cartoon-videos detected as a ‘cartoon’ and no misclassification of other genres as cartoons was achieved using a combined use of audio and video features. Neither audio nor video features alone achieved this result. In summary, a robust detection of cartoons is possible with essentially no other genre classified as cartoon. This is important in application that requires that users (children) are not provided with broadcast other than cartoons.

With our current non-optimized software system we achieved on an AMD Athlon XP1600+, 1.41 GHz a run-time performance of approximately 1 min for classification of 1 min of video. We took great care to perform experiments on a large and balanced database of video from real, compressed broadcast material.

6. REFERENCES

[1] T.I. Ianeva, A.P. de Vries and H. Röhrig, “Detecting cartoons: A case study in video-genre classification”, Proceedings ICME Multimedia and Expo, volume 1, pp. 449-452, 2003.

[2] M. Roach, J.S. Mason and M. Pawlewski, “Motion-based classification of cartoons”, International Symposium on Intelligent Multimedia, pp. 146 – 149, 2001.

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

[4] L.Q. Xu and Y. Li, “Video classification using spatial-temporal features and pca”, Proceedings ICME Multimedia and Expo, volume 3, pp. 485-8, 2003.

[5] M. Roach, J. Mason and L.-Q. Xu, “Video genre verification using both acoustic and visual modes”, International Workshop on Multimedia Signal Processing, pp. 157 – 160, 2002.

[6] M. Roach, J. Mason, N. Evans and L.-Q. Xu, “Recent trends in video analysis: a taxonomy of video classification problems”, Proceedings Internet and Multimedia Systems and Applications, pp. 348 – 354, 2002.

[7] M. Roach, J. Mason and M. Pawlewski, “Video genre classification using dynamics”, Proceedings Acoustics, Speech, and Signal Processing, volume 3, pp.1557 – 1560, 2001.

[8] B. T. Truong, S. Venkatesh and C. Dorai, “Automatic genre identification for content-based video categorization", Proceedings 15th International Conference on Pattern Recognition, volume 4, pp. 230-233, 2000.

[9] S. Fischer, R. Lienhart and W. Effelsberg, “Automatic recognition of film genres”, third ACM International Multimedia Conference and Exhibition, pp. 295-304, 1995.

[10] J. Canny, “A computational Approach for Edge Detection”, IEEE Trans. Pattern Anal. Machine Intell., volume 8, no. 6 pp. 679-698, 1986.

[11] H. Krambeck, R. Glasberg and T. Sikora, “Development of an Analyzer for Video-Genre-Classification of mpeg-2 streams in digital video broadcasting”, TU-Berlin, student research project, 2004.

[12] T. Sikora, P. Salembier and B.S. Manjunath, “Introduction to MPEG-7: Multimedia Content Description Interface”, John Wiley LTD, ISBN 0471486787, 2002.

[13] Z. Liu, Y. Wang, and T. Chen, “Audio Feature Extraction and Analysis for Scene Segmentation and Classification”, Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology, pp. 61-80, 1998.

[14] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi and T. Sorsa “Computational Auditoy Scene Recognition”, Proceeding ICASSP International Conference on Acoustics, Speech and Signal Processing, volume 2, pp.1941 – 1944, 2002.