Real-Time Detection of Sport in MPEG-2 Sequences using High-Level AV-Descriptors and SVM

Ronald Glasberg[1], Sebastian Schmiedeke[2], Hüseyin Oguz[3], Pascal Kelm[4] and Thomas Sikora[5]

Communication Systems Group

Technische Universität Berlin, Germany

{[1]glasberg [2]schmiedeke [3]oguz [4]kelm [5]sikora}@nue.tu-berlin.de

***Abstract* — We present a new approach for classifying mpeg-2 video sequences as 'sport' or 'non-sport' by analyzing new high-level audiovisual 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 video, news and sports are studied. Such applications have also been discussed in the context of MPEG-7 [1]. In our method the extracted features are logically combined by a support vector machine [2] to produce a reliable detection. The results demonstrate a high identification rate of 98.5% based on a large balanced database of 100 representative video sequences gathered from free digital TV-broadcasting and world wide web.**

I. Introduction

With the advent of digital TV-broadcasting, e.g. DVB and IPTV presenting more than hundreds of channels at a time, the need for a user-friendly TV-program selection is growing. Unlike the present situation, a new system should enable users to access programs clustered by genres according to Fig. 1.

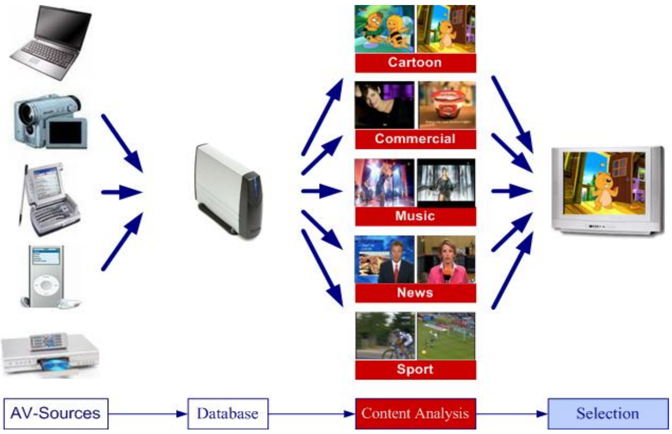

Fig. 1. Concept of a Video-Genre-Classification System

Our interest is the real-time detection of sport videos in broadcasting TV material.

II. Related Work

Sport is a diverse genre, consisting of sub-genres like basketball, football, motor, soccer, tennis etc. In order to reliably detect sport, research has focused on two approaches listed in Table I.

The first approach 'intra-sport' classification tries to detect different sport types [3]-[5] or events [6]-[8] within the sport domain; whereas the second approach 'inter-genre' classifies cartoon, commercial, music, news and sport as part of the multi-category problem [9]-[11].

TABLE I
SELECTED PUBLICATIONS RELATED TO SPORTS DETECTION

	Approach, Publ.	Genres	Database & Cliplength	Video-features	Audio-features	Classifier	Results ($T_{decision}$)
Intra-Sport-Detection	[3]	BASK, BOX, DIV, FOOT, GOLF, TENN, VOLL	27 Clips à 8-9 min	Color correlogram (spatial correlation)	/	SVM (RBF)	CA = 91%
	[4]	BASK, SOCC, TENN vs. NEWS	4 Clips à 4 h	Camera Motion, Motion (intensity, direction), Dominant Color (global, local)	MFCC	Pseudo-2D-HMM	CA = 100% (96 s)
	[5]	BASE, BASK, FOOT, SOCC, TENN, VOLL	600 Clips à 3-10 min	Shot Length, Cut, Camera Motion, Face ratio, Brightness, Color (Diff, Entropy)	/	SVM Binary-Tree (C4.5)	CA = 95%
	[6]	Fieldsport-Events	100 h	Hard Cut, Close Up, Crowd Image, Scoreboard, Motion Activity, Field Lines	Speech Band Audio Activity	SVM	Max. ERR = 97% (25 s)
	[7]	Shots of SOCC	20 Clips à 2-10 min	Color (Field Area), Motion Intensity	/	HMMs Hierarchy	CA = 83%
	[8]	Shots of SOCC	not mentioned	19 features like Gras Color Distrubition, Edge Distribution, Shot Length ...	/	SVM	CA = 92%
Inter-Genre-Detection	[9]	CAR, COM, MUS, NEWS, SPO	5x 1 h à 5 min	Scalable Color, Color Layout, Homogeneous Texture	MFCC	GMM	CA = 87%
	[10]	CAR, COM, MUS, NEWS, SPO	5x 1 h à 5 min	Motion Activity	MFCC	GMM	ERR = 10% (20 s)
	[11]	CAR, NEWS, SPO	18 Clips à 50 s	Motion (Camera, Object)	/	GMM	EER = 6% (30 s)
	[our]	SPO vs. NON-SPO	5x 20 Clips à 3 min	Sportfield Color, StaticArea, Scoreboard, Logo Recognition	Noise Level	SVM (RBF)	CA = 99% (20 s)

BASE = Baseball, **BASK** = Basketball, **BOX**- Boxing, **CYCL** = Cycling, **DIV** = Diving, **FOOT** = American Football, **ICEH** = Icehockey, **SOCC** = Soccer, **SWIM** = Swiming, **TENN** = Tennis, **VOLL** = Volleyball, **YACH** = Yachting; **CAR** = Cartoon, **COM** = Commercial, **MUS** = Music video, **NEWS** = News video, **SPO** = Sport video

Watcharapinchai et al. [3] analyze the sport types basketball, boxing, diving, football, golf, tennis and volleyball by using a low-level feature 'color spatial correlation' and two alternative classifiers: a neural network with principal component analysis (PCA) and a support vector machine (SVM) with radial basic function (RBF) kernel. The best result was achieved by the SVM with classification accuracy CA = 91%.

Wang et al. [4] use low-level audiovisual features such as camera motion, motion (intensity and direction), dominant color (global and local) and 13 Mel-frequency cepstral coefficients (MFCC). To recognize the temporal pattern of a sport sequence a Pseudo-2D-Hidden Markov model (HMM) was implemented. The achieved result on a decision window of T_{Decision}=96s is CA = 100%.

Yuan et al. [5] analyze similar to [3] the sport types baseball, basketball, football, soccer, tennis and volleyball by extracting low-level features grouped in temporal (shot length, cut percentage, color difference, camera motion) and spatial dimension (face frame ratio, brightness, color entropy). With a hierarchical ontology (binary tree with cross-validated SVM) they achieve a CA = 95%. An extension of these publications related to ‘intra-sport’ classification is [6]-[8]. They detect specific events within the sport domain, like play/ break scenes in soccer.

The second approach, the ‘inter-genre’ classification of cartoon, commercial, music, news and sport as part of the multi-category problem are presented in [9]-[11]. They use low-level audiovisual features like color layout, motion activity, camera motion, scalable color, homogeneous texture and MFCC to solve the problem.

The mentioned publications for sport detection use mainly low-level features and achieve accuracies up to 95% with their own databases and T_{Decision}, which makes a comparison very difficult.

In our ‘inter-genre’ approach we involve mainly high-level descriptors combined with a low-level motion feature and achieve a higher CA = 98.5% on a representative database.

III. Classification Process

In a first step the audiovisual data is demultiplexed from mpeg-2 streams in a video and an audio sequence.

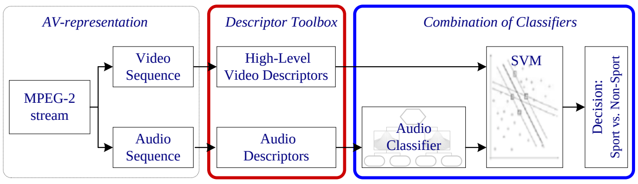

Fig. 2. Process involved in computing the feature vectors

We consider all consecutive frames as well as compressed stream information and provide them for further processing.

A. Video Descriptors

Sport videos, mainly diverse, have some features in common, like characteristic sport fields, specific logos, scoreboards, motion, spectators, sports equipment etc. We implemented dedicated features for classifying videos as ‘sport‘ or ‘non-sport’. In following the most relevant features are described.

1) *Scoreboard Descriptor:* To encode MPEG-2 files each frame in the original video is divided into 8x8 blocks and then the discrete cosine transform (DCT) is applied to these individual blocks. The resulting I frames include these DCT coefficients, while the P frames additionally include motion information. The first coefficient of the DCT of a block is named DC-coefficient and it is proportional to the average intensity of that block, the high-order coefficients named AC-coefficients inform about the block’s angularity. The scoreboard descriptor in Fig. 3 intents to detect static text areas appearing usually in the four corners M_1-M_4 of n successive frames. Therefore we extract the sum of AC-coefficients in these corners in a first step. For the corners with a sum higher than a threshold Th_1 we also analyze the motion vectors of the consecutive P-frames. The areas with high number of edges and nearly static (no motion activity) are highlighted as potential scoreboards, if the number of these areas exceeds the threshold Th_2.

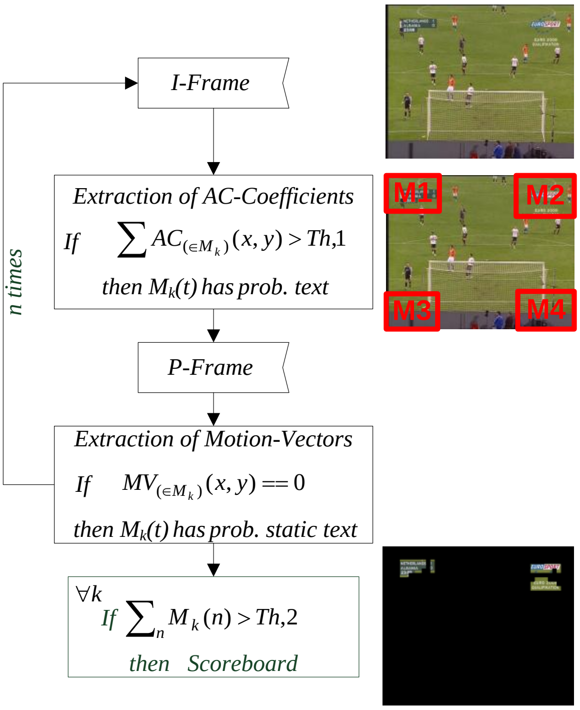
I-Frame
Extraction of AC-Coefficients
If $\sum AC_{(\in M_k)}(x,y) > Th,1$
then $M_k(t)$ has prob. text
P-Frame
Extraction of Motion-Vectors
If $MV_{(\in M_k)}(x,y) == 0$
then $M_k(t)$ has prob. static text
$\forall k$
If $\sum_n M_k(n) > Th,2$
then Scoreboard
n times
M1
M2
M3
M4

Fig. 3. Block diagram of the scoreboard descriptor

2) *Sport Field Descriptor*: The descriptor in Fig. 4 is based on the idea, that if a sport field is visible, the field color is the predominant color of a frame and remains unchanged over a certain period of time. We implemented the idea by determining the most dominant color $Centre_{\text{Lab}}$ of an I-frame. This detected centre is tracked over n consecutive frames and the number of frames is counted, in which the centre remains at the same position within the threshold Th_1. A sport field is assumed, if this counter $N_{\text{SameColor}}$ exceeds the threshold Th_2. For a sequence of frames fulfilling this condition, we execute a simple template matching. The detected field color $Centre_{Lab}$ will be assigned to one of our predefined templates $Centre_{FieldTemplates}$ by using the Euclidean distance. These templates include field sports like soccer, football, hockey and individual sports like tennis on different grounds.

3) Logo Recognition: The descriptor in Fig. 5 uses the grayscale information from the four corners M_1-M_4 of n successive I frames. Pixels of these corners with constant values over a period of time suggest the appearance of a logo and increase the counter of the corresponding M_{k} matrices. After n frames we map the matrices M_1 - M_4 to horizontal and vertical vectors $S_{hor,k}$ and $S_{ver,k}$ via summation. Finally these vectors are compared with predefined logo vectors through a cross-correlation function. A logo is recognized, if the correlation coefficients exceed the threshold Th_2.

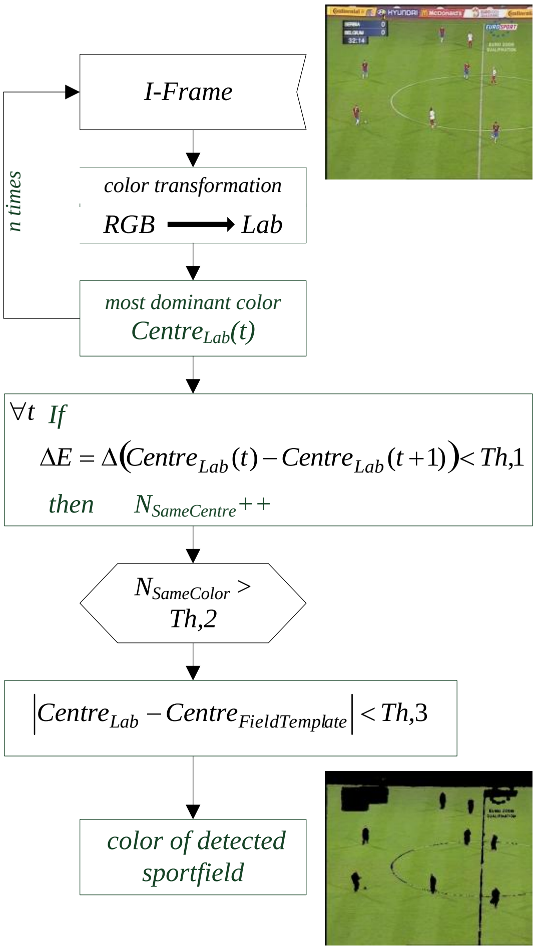
I-Frame
n times
color transformation
RGB → Lab
most dominant color
$Centre_{Lab}(t)$
$\forall t$ If
$\Delta E = \Delta(Centre_{Lab}(t) - Centre_{Lab}(t+1)) < Th,1$
then $N_{SameCentre}{+}{+}$
$N_{SameColor} >$ Th,2
$|Centre_{Lab} - Centre_{FieldTemplate}| < Th,3$
color of detected sportfield

Fig. 4. Block diagram of the sport field descriptor

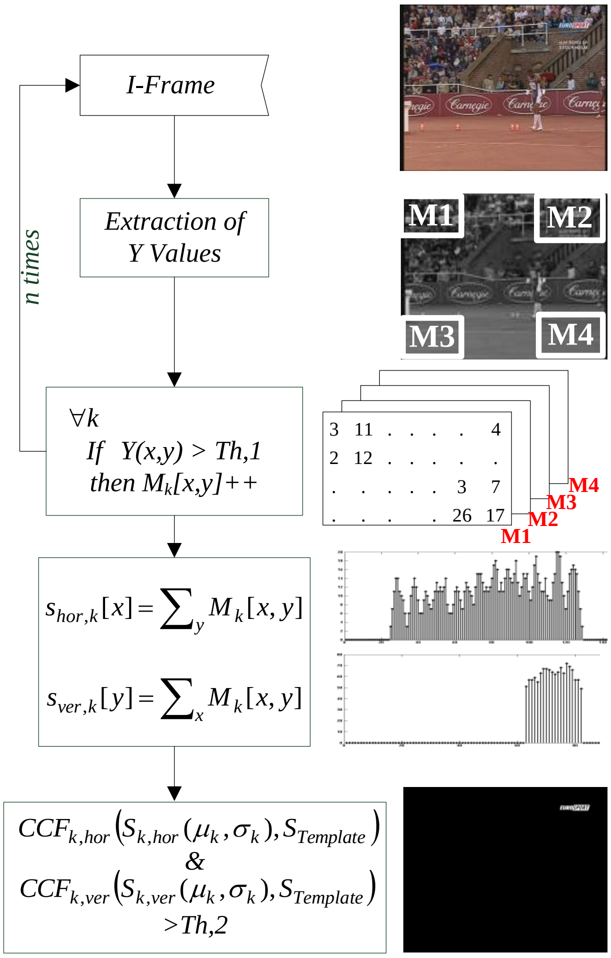
I-Frame
Extraction of Y Values
n times
$\forall k$
If Y(x,y) > Th,1
then $M_k[x,y]{+}{+}$
M1
M2
M3
M4
3 11 4
2 12 . . . . .
. . . . . 3 7
. . . . 26 17
$s_{hor,k}[x] = \sum_y M_k[x,y]$
$s_{ver,k}[y] = \sum_x M_k[x,y]$
$CCF_{k,hor}(S_{k,hor}(\mu_k,\sigma_k), S_{Template})$
&
$CCF_{k,ver}(S_{k,ver}(\mu_k,\sigma_k), S_{Template})$
$> Th,2$

Fig. 5. Block diagram of the logo recognition descriptor

B. Audio Descriptors

Several low-level audio-descriptors [12] such as Harmonicity, High Zero Crossing Rate, Low Energy Frame Rate and Spectrum Spread Modulation are used to distinguish between the audio categories ‘music’, ‘noise’, ‘silence’ and ‘speech’ according to Fig. 6. For our sport detection we are only interested in the ‘noise’ level information.

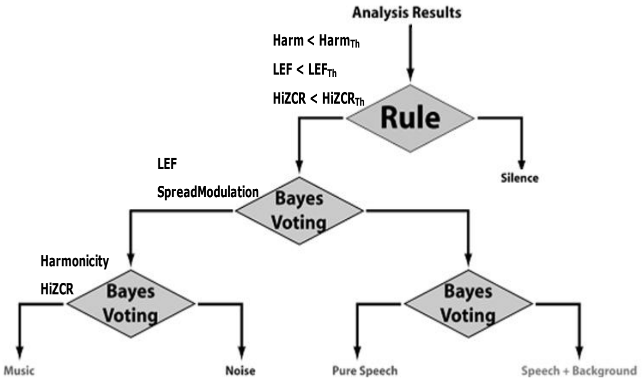

Fig. 6. Structure of the audio classifier

C. Classification

We distinguish between ‘sport’ and ‘non-sport’ videos by using the three presented high-level descriptors, a low-level visual descriptor ‘static area’ and the audio ‘noise’ level.

These five audiovisual features are used as input for different classifiers. We tested the following methods:

- A Bayesian classifier with a single Gaussian per feature and priors calculated according to distribution of training sets
- A decision tree according to the ID3 algorithm [13], the real-valued features are binned into specific intervals and handled as nominal attributes,
- A multi-layer perceptron (MLP) with a hidden layer containing 50 neurons and non-linear sigmoid transfer functions
- A support vector machine [2] with C=8 as cost factor and radial basic function kernel with γ=1/8 as parameter specifying the width of Gaussian

We achieve the best classification accuracy with a SVM and a decision window of T_{Decision}=20 sec, the details are shown in Table II.

TABLE II
ACCURACY OF DIFFERENT CLASSIFIERS

method	precision	recall	CA
Bayes	70.4%	91.3%	90.5%
Decision tree	100%	92.1%	98.4%
MLP	100%	67.9%	93.5%
SVM	97.5%	95.1%	98.5%

The SVM assigns a video sequence by the following rule

$$f(\vec{x}) = sign\left(\sum_{sv} \lambda_{sv} z_{sv} \cdot \kappa\left\langle \vec{x}_{sv}, \vec{x}\right\rangle + b\right)$$

wiht the support vectors $\vec{x}_{sv}$ and Lagrange factor λ_{sv}, class labels z_{sv} and RBF kernel $\kappa\langle u,v\rangle = e^{-\gamma\cdot|u-v|^2}$.

The structure of used SVM is shown in Fig. 7.

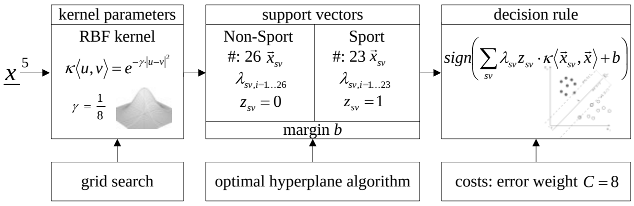

Fig. 7. Structure of the implemented SVM

IV. EXPERIMENTS

The experiments were carried out on a database of 100 representative mpeg-2 video sequences (50 sequences for training and 50 as testing-data), in total of 300 min of recordings; 20 'sport' videos and 4•20 'non-sport' videos (cartoon, commercial, music and news) of 3 minutes' each gathered from popular TV (ARD, BBC, EuroSport etc.) and world wide web. As each descriptor takes different processing time T_{Proc} to extract the desired feature, the classifier has to wait in our case T_{Decision} = 20 sec until all features are available.

A. *Experimental Results*

1) The performance of the sport field descriptor in soccer, football, hockey and tennis on different grounds is shown in Table III.

TABLE III
ACCURACY OF THE SPORT FIELD DESCRIPTOR

tested sequences	# decisions	# correct decisions	% correct decisions
Cartoon	1236	1066	86.3%
Commercial	1293	1231	95.2%
Music	1346	1343	99.8%
News	2112	1968	93.1%
Sport	2146	1995	93.0%

The total accuracy is 93.6% within T_{Proc} =5 I-frames. The misdetections are caused by static and homogenous backgrounds. Fig. 8 illustrates that cartoon and news sequences were assigned as soccer (20) and as tennis sport field (30).

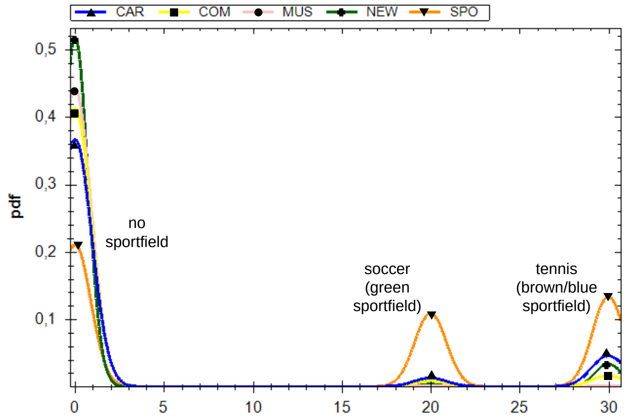

Fig. 8. Pdf of sport field descriptor

2) The result of the logo recognition is shown in Table IV. The descriptor achieves an accuracy of 96.1% within T_{Proc} = 35 I-frames.

TABLE IV
ACCURACY OF LOGO RECOGNITION

tested sequences	# decisions	# correct decisions	% correct decisions
Cartoon	169	168	99.4%
Commercial	179	176	98.3%
Music	184	184	100%
News	290	279	96.2%
Sport	291	263	90.4%

The best result is achieved by music videos. The misdetections in sport are caused by the variation of certain logos adapted to different sport events. Fig. 9 illustrates the different logo templates in increments of 10.

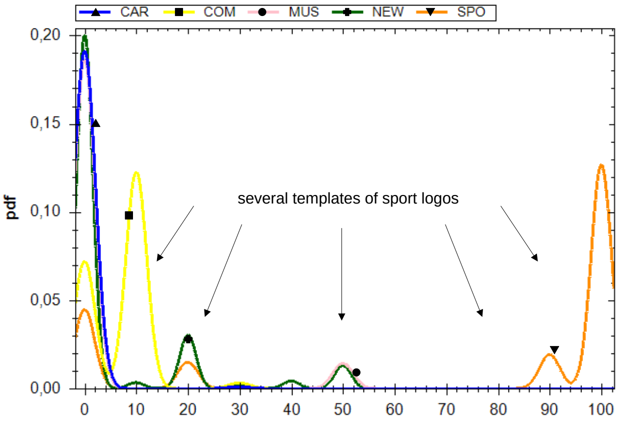

Fig. 9. Pdf of logo recognition descriptor

3) The results of the scoreboard descriptor are shown in Table V.

TABLE V
ACCURACY OF SCOREBOARD DESCRIPTOR

tested sequences	# decisions	# correct decisions	% correct decisions
Cartoon	614	529	86.2%
Commercial	628	556	88.5%
Music	667	486	72.9%
News	1052	725	68.9%
Sport	1066	876	82.2%

The accuracy of 78.8% within T_{Proc} =10 I-frames is encouraging and will be improved in future research. The main challenge is that the descriptor cannot distinguish between general text fields in the corners of a video sequence and specific scoreboards in sport (Fig. 10).

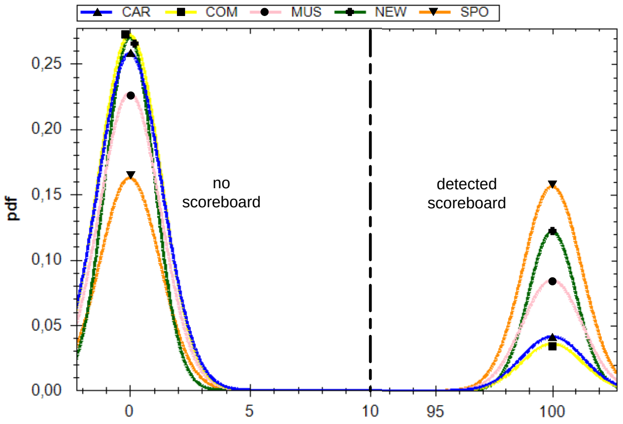

Fig. 10. Pdf of scoreboard descriptor

4) The classification performance of the SVM sport video detector is presented in Table VI.

TABLE VI
CLASSIFICATION MATRIX AND ACCURACY

true \ pred	non-sport	sport	recall
Cartoon	65	0	100%
Commercial	80	1	98.8%
Music	70	0	100%
News	105	1	99.1%
Sport	4	77	95.1%
precision	98.8%	97.5%	**CA = 98.5%**

From the cartoon, commercial, music and news videos, more than 99% decision windows (320 from 322) were correctly detected as ‘non-sport’. The two misclassifications were caused by sport scenes in a commercial and news video.

In ‘sport’ more than 95% of the decision windows (77 from 81) were correctly classified.

V. SUMMARY & CONCLUSION

In this paper we presented a new approach to detect sport videos in a database consisting of cartoon, commercial, music, news and sport videos. We started with high-level audiovisual descriptors and a support vector machine to combine the results, deriving a probability, for a video sequence being ‘sport’ or ‘non-sport’.

In comparison to recent inter-genre related work [9]-[11] using mainly low-level descriptors, our approach achieves an accuracy of 98.5% with high-level descriptors. The classification accuracy could be improved increasing the duration of the decision window.

With our current non optimized software system we achieved on an AMD Athlon64 X2 5000+, 2.61 GHz a run-time performance of approximately 1 min for classification for 1 min of video. Future research will be done to extend and improve our descriptor toolbox [14] of high-level descriptors.

REFERENCES

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

[2] C. Cortes and V. Vapnik, Machine Learning: Support-vector networks, Springer Netherlands, Vol. 20, No. 3, pp. 273-297, Sep. 1995.

[3] N. Watcharapinchai, S. Aramvith, S. Siddhichai, S. Marukatat, “A discriminant approach to sports video classification”, International Conference on Communications and Information Technologies, ISCIT '07, pp. 557 – 561, 17-19 Oct. 2007.

[4] J. Wang, C. Xu, E. Chng, “Automatic Sports Video Genre Classification using Pseudo-2D-HMM”, 18th International Conference on Pattern Recognition, ICPR 2006. Vol. 4, pp.778 – 781, 2006.

[5] X. Yuan, W. Lai, T. Mei, X.-S. Hua, X.-Q. Wu, S. Li, „Automatic Video Genre Categorization using Hierarchical SVM”,
IEEE International Conference on Image Processing, pp. 2905 – 2908, 8-11 Oct. 2006.

[6] D.A. Sadlier and N. O'Connor, “Event Detection in Field Sports Video Using Audio-Visual Features and a Support Vector Machine”, Transactions on Circuits and Systems for Video Technology, Vol.15, 2005.

[7] F. Wang, Y.F. Ma, H.J. Zhang and J.T. Li, „A Generic Framework for Semantic Sports Video Analysis Using Dynamic Bayesian Networks”, Proceedings of the 11th International Multimedia Modeling Conference (MMM), 2005.

[8] Y.-H. Zhou, Y.-D. Cao, L.-F. Zhang, H.-X, Zhang, „An SVM-based soccer video shot classification”, International Conference on Machine Learning and Cybernetics, Vol. 9, pp. 5398 – 5403, 18-21 Aug. 2005.

[9] L.Q. Xu and Y. Li, “Video Classification Using Spatial-Temporal Features And PCA,” Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), 2003.

[10] M. Roach, J.S. Mason and L.Q. Xu, “Video Genre Verification using both Acoustic and Visual Modes,” Proceedings of the IEEE Workshop on Multimedia Signal Processing, 2002.

[11] M. Roach, J.S. Mason and M. Pawlewski, “Video Genre Classification Using Dynamics,” Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2001.

[12] H.G. Kim, N. Moreau and Thomas Sikora, Introduction to MPEG-7 Audio: Content Indexing and Retrieval, Wiley & Sons, ISBN-10: 047009334.

[13] J.R. Quinlan, “Decision trees and decision-making”, IEEE Transactions on Systems, Man and Cybernetics, vol. 20, pp. 339 - 346, 1990.

[14] R. Glasberg, S. Schmiedeke, M. Mocigemba and T. Sikora, “Real-Time Approaches for Video-Genre-Classification using new High-Level Descriptors and a Set of Classifiers”, Proceedings of the Second IEEE International Conference on Semantic Computing, 2008