HIGHLIGHTS DETECTION IN SPORTS VIDEOS BASED ON AUDIO ANALYSIS[1]

Hadi Harb, Liming Chen
LIRIS CNRS FRE 2672
Ecole Centrale de Lyon
Dépt. Mathématiques Informatiques, 36 avenue Guy de Collongue
69134 Ecully, France
{Hadi.Harb, Liming.Chen}@ec-lyon.fr

ABSTRACT

While it is very hard to achieve automatic sports competition key moments detection only based on visual analysis, we propose in this paper automatic highlights detection based on an audio classifier. The audio classifier is based on a new modeling technique of the audio spectrum called Piecewise Gaussian Modeling (PGM) and Neural Networks. The proposed approach was evaluated on soccer and tennis videos, though our technique has no restriction on the sports' type. It is shown that audio-based highlights detection can be effective for tennis segmentation since 97.5% of end-of-serves were correctly classified. Goals can be detected in soccer videos using audio analysis as well. An intelligent sports-videos player is proposed based on the audio analysis permitting the user to navigate through key moments in a sports video.

1. INTRODUCTION

Content-based video indexing has achieved great strides in the last decade. However, developing general purpose algorithms capable of analyzing all types of videos is still a very difficult problem. In order to be as efficient as possible, video indexing techniques must take into account as much as possible a priori knowledge. In this paper, we investigate sports video analysis from audio stream. Sports videos are very common, especially in the entertainment market. Hundreds of millions of people are used to watch FIFA World Cup for example. The same can be seen in other types of sports such as American football, or tennis, etc. As sports program are often live broadcasted and the amount of sports videos existing today is huge, the need for automatic systems that help content-based indexing and/or intelligent navigation for sports videos is clearly becoming crucial [18].

Special events or highlights detection in sports videos consists of detecting key moments corresponding to semantically important actions such as goals or goal attempts in soccer. Special events can be the basis of an intelligent sports-video player. The user can navigate into a sports video based on the "action indicator".

In this paper we investigate the use of the audio stream analysis for special events' detection. The audio stream in sports videos carries a great amount of semantic information. For example, when listening to a soccer match the listener can easily and efficiently detect important moments in the game with a minimum effort. Although the experiments in this paper were carried out on two types of sports which are tennis and soccer, applying the same technique to other types of sports videos is straightforward.

We apply an audio classifier aiming at classifying the audio stream in a sports video into action, and no-action [24]. The definition of the two classes can differ from one type of sports to another. A training phase where the classes must be defined is always needed. A manual classification of the training data is essential for the training phase. However, the audio classifier presented in this paper does not require a big amount of training data (it generally needs 20 seconds for each class) making the training phase a simple problem.

The audio classifier is based on a new modeling technique of the audio spectrum, called Piecewise Gaussian Modeling (PGM), and a Neural Network as a classifier.

The rest of the paper is organized as follows: Section 2 gives a resume of the related work in the field of sports videos analysis, Section3 describes our audio classifier, Sections 4 and 5 provide the details of the application of the audio classifier to highlights detection in tennis and soccer. We conclude in Section6.

[1] This work has been partially supported by the project Cyrano within the framework of French national research network on telecommunication RNRT.

2. RELATED WORK

Recently considerable work has been done in the context of sports videos analysis. Systems that resume, segment, describe sports videos have been proposed. However, the majority of these systems consider some domain knowledge, so they are suitable for specific types of sports. This is normal since each sport has its specific rules and hence an analysis algorithm can rely on such a priori knowledge for better performance.

We can group algorithms for sports videos analysis into two categories: 1- Algorithms aimed at detecting highlights in a sports video, or special events. 2- Algorithms aiming at structuring or describing a sports video.

Highlights detection consists of detecting some special events (semantic events) in a sports video. The work presented in this paper belongs to this category. [10] uses template matching of Fast Fourier Transform (FFT) features for audio-based impact recognition in tennis. [20] also uses FFT template matching for wordspotting in the audio stream of sports videos, and the energy envelope of the audio signal is used for cheers detection for American football. In [16] the energy of frequencies higher than 700 Hz combined with cut rate and motion activity were used for action detection in soccer. The energy level of audio signal was also used by [19] for special events detection in soccer videos. Color-based features were used in [15] to detect important segments in sports videos (such as plays in soccer) in the goal of enabling a content-based adaptive streaming of sports videos. Visual features and Hidden Markov Models (HMM) were used in [3] for the classification of soccer videos into play/out-of-play. Camera movements were used in [14] to detect events such as ball serves in tennis and volleyball. Camera movements and player/ball tracking were used by [2] for highlights detection in soccer videos. An attempt to detect shots at goals in soccer videos using motion, ball, players, and lines tracking was made by [6]. Hidden Markov Models (HMM) and camera movements features were proposed by [11]for highlights detection in soccer. In [4] the highlights are supposed to be extracted manually and a system that learns the user's preferences automatically was proposed. An audio-based approach is presented in [1] for highlights detection in baseball videos.

Other types of analysis can be carried out on sports videos. [17] classifies sports videos into football or basketball. [12] uses image analysis techniques for the tracking of players in soccer videos in order to describe a game. Court line detection and player tracking were used in [7] to describe tennis videos. [8] uses color-based features to track players and the ball in tennis videos in the aim of providing statistics of the players' strategies in the game. A system that indexes in real time tennis videos using player and ball tracking is presented in [9]. A rule-based approach based on audio and visual features is described in [5] in the aim of structuring basketball videos; whistles, speech, and noise constitute the audio classes that are recognized.

As we can see, the audio stream has got relatively little attention by the researchers when designing systems for sports videos analysis. Until now, relatively simple solutions have been proposed for audio analysis.

3. AUDIO CLASSIFICATION

Special events detection is a two class classification problem. The definition of the audio classes is related to the problem in hand and to the user's needs. For instance, in the case of a tennis competition, one class is the applause and the other one is the no-applause.

In our approach, the audio stream is extracted from the video stream and the signal is down-sampled to 8 KHz. The signal is windowed with a Hamming window of 31.25 ms width and 21.25 ms overlap. The Fast Fourier Transform (FFT) is applied on each window (or audio frame), meaning that a spectral vector containing the magnitudes of FFT coefficients is extracted every 10 ms. The spectrum is then filtered using a filter-bank (19 filters) distributed based on the MEL scale. Thus, every 10 ms one vector containing 19 coefficients, the Mel Frequency Spectral Coefficients (MFSC), is obtained.

MFSC features do not take into account any information about the audio classes, making them general. The use of general audio features, such as MFSC, is extremely important in this case of audio classification since the audio classes are not known *a priori* and their definition can change based on the application and/or the user's needs.

3.1. Piecewise Gaussian Modeling

Relatively long term audio features are needed to easily perceive our classes. For instance, the excitation in the commentator's voice cannot be classified if we only listen to 10 ms of audio. We propose to classify the audio signal based on features extracted from time windows of 1s.
Therefore, the signal is segmented into non-overlapped windows of 1 second which constitute the basic frames for the classification.

The MFSC vectors in each window are modeled by one Gaussian model expressed by a mean vector and a variance one. That is to say, one audio segment of duration M seconds is modeled by M Gaussian models: $\mathrm{N}_0(\mu_0,\nu_0),\mathrm{N}_1(\mu_1,\nu_1),...\mathrm{N}_N(\mu_M,\nu_M)$ Figure 1. This is what we call Piecewise Gaussian Modeling (PGM) of the MFSC vectors. The term "Piecewise" is used since a set of spectral vectors contained in an audio segment are not modeled as a whole using a set of Gaussian Models. Instead, they are modeled by one Gaussian Model every T seconds (lets say 1 s) step, hence preserving local time information. That is, the neighborhood of the consecutive short term spectral vectors contained in a long term window is preserved by the statistics. In each frequency channel, the distribution of the values of the magnitude is expressed within one second time window by a mean and a variance. The difference between PGM and Gaussian Mixture Models (GMM) is that in the former the relation between consecutive spectral vectors is modeled and the time information is preserved while in the latter the spectral vectors are modeled with no time information and the relation between spectral vectors is not taken into account. This means that PGM preserves the dynamic features of the spectrum while conventional GMM doesn't. Dynamic features are shown to be important for speech recognition and speaker recognition applications where the derivatives of spectral/cepstral vectors is employed to model the dynamics [21][23].

The concatenation of the mean and the variance values for each 1 second window constitutes the feature vector Figure 2. Therefore, the Neural Network will be trained to classify a distribution of a set of consecutive spectral vectors described by its Gaussian parameters (Mean, Variance).

On the other hand, this feature vector is aimed at capturing the timbre of the audio signal since it includes the distribution of the energy within each frequency channel.

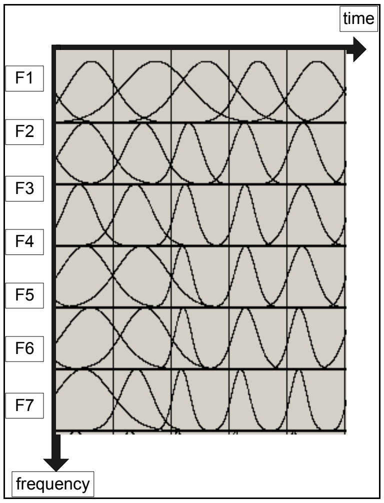

Figure 1 In each frequency band and for each time window (1 s) one Gaussian model of the spectral magnitude is obtained.

$$\begin{array}{ccccccccc}
\begin{bmatrix}X1\\X2\\X3\\.\\.\\.\\Xk\end{bmatrix} & \begin{bmatrix}X1\\X2\\X3\\ \\ \\ \\Xk\end{bmatrix} & \begin{bmatrix}X1\\X2\\X3\\ \\ \\ \\Xk\end{bmatrix} & \begin{bmatrix}X1\\X2\\X3\\ \\ \\ \\Xk\end{bmatrix} & \begin{bmatrix}X1\\X2\\X3\\ \\ \\ \\Xk\end{bmatrix} & \begin{bmatrix}X1\\X2\\X3\\ \\ \\ \\Xk\end{bmatrix} & \Rightarrow & \begin{bmatrix}\mu 1\\\mu 2\\\mu 3\\ \\ \\ \\\mu k\end{bmatrix}, & \begin{bmatrix}\nu 1\\\nu 2\\\nu 3\\ \\ \\ \\\nu k\end{bmatrix} \\
t & ,t+1, & t+2, & & & t+N & & &
\end{array}$$

Figure 2 The spectral Vectors in each time window (1 s) are modeled by one Mean Vector and One Variance Vector.

3.2. Neural Network as a classifier

The feature vector obtained for each time window can be classified by any classifier. However, we use Multi Layer Perceptron (MLP) as a classifier for the following reasons:

1- the MLP can generally be effective when the decision boundary between the classes has a complex shape

2- The classification using the MLP is not computationally expensive, making it suitable for applications where the speed of the classification is important.

3- Once a MLP is trained the transfer of the knowledge that permits the classification can be done by transferring the synaptic weights and the architecture which implies a compact representation of the classifier.

3.3. Architecture of the audio classifier

The audio classifier that we use is based on three main modules: audio signal processing module, statistics and normalization module, and neural network module Figure 3.

Signal processing module: this module is responsible for FFT calculation, MEL filtering and it provides a set of MFSC vectors in each 1 second window

Statistics and normalization module: this module computes the statistics of MFSC vectors, that is the mean and variance vectors. Also this module normalizes the mean and the variance vectors by their respective maximum. This normalization is extremely important for the classifier to learn the relation between the frequencies and their distribution with no relation to the energy, leading to a robustness to channel and loudness changes. Moreover, normalizing the values of the feature vector is important when using the MLP as a classifier since high values in the feature vector may lead the MLP to saturation [22].

Neural Network module: this module has the normalized feature vector as an input. We use an MLP with one hidden layer and the error back-propagation algorithm as a training algorithm.

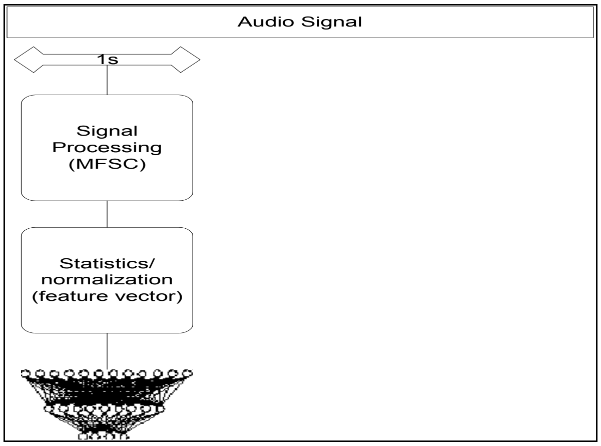

Figure 3 The architecture of the audio classifier.

3.4. The limits of HMM

Hidden Markov Models (HMM) have been used successfully for the problems of audio classification and recognition, namely speech recognition. HMM have the ability to model the temporal relation between consecutive feature vectors. A HMM contains a set of states and each state have a Probability Distribution Function *p.d.f.* that governs the probabilities of emitting one particular feature vector. The transition between states is governed by a probability transition matrix.

To be efficient, the states of a HMM must describe a natural stable states in the feature space. Also, the transition between states must have physical sense in order to correctly model the time information. In the case of speech recognition, the states relate to phonemes and the transition between states follows rules related to the spoken language and its structure. Hence, HMM are efficient for speech recognition.

However, in this case of highlights detection in sports videos, clear states do not exist. Therefore, two basic states that can be defined are: action and no-action. The transition between these two states does not have clear physical sense since it is closely related to the audio sequence, the sports game for instance. An excited soccer game will probably have numerous transitions between the two sates, other than for a calm game the probability of staying in the no-action state will become greater. Therefore, a great amount of training data will be needed to efficiently cover the variability of sports games making retraining and changing the classes' definitions a difficult process. This implies that the use of HMM to this classification problem in real world conditions is not motivated.

4. TENNIS AUDIO ANALYSIS

Generally speaking, the aim of audio classification for tennis videos analysis is twofold: 1- providing a new method enabling content-based navigation through a tennis video based on the probability of "special events", 2- aiding image-based video segmentation of a tennis competition.

The task in the tennis segmentation problem is to segment a tennis competition into serves, games, sets, etc. Knowing that generally at the end of each serve we have cheers, audio analysis (applause detection) can provide a helpful cue to a tennis video segmentation algorithm. Generally the spectators' applause is related to the time when a semantically special event occurs. Moreover, the time of the applause is an indicator to importance of the special event. For instance, when the applause's duration is 10 seconds, one can expect that an important special event has occurred.

In our first experiment, we have applied our audio classification engine for applause recognition in tennis videos. Once recognized, the applause moments within the audio stream, a tennis video player that indicates the time of the applause to the user can be efficient for an intelligent navigation into a tennis game Figure 4. Automatic tennis video summarization can also be performed based on the applause information. Such an approach for tennis video navigation or summarization would probably fail if the spectators do not appreciate the special events. However, the experience of a user using such a tennis video player will be strongly related to the spectators' experience.

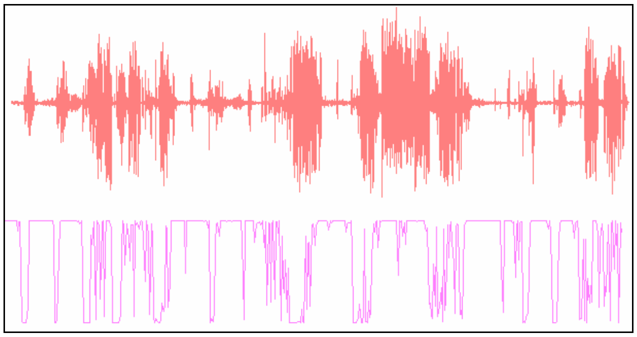

Figure 4 Playing sports videos using the audio-based highlights detection. At the top the audio signal is shown, and at the bottom the probability of highlights.

Hence the user can navigate through the video based on highlights probability.

4.1. Experiments

Three tennis videos from the Australian open 2002 were selected for the experimentation. 10 minutes from each match were used. The audio classifier was trained on 20 seconds of special events (applause) and 20 seconds of non-special events (speech and silence). The training data was extracted from one of the three tennis videos. The test data was manually labeled as special event or non-special event. Generally special events were after an out of play. The test data contained 42 special events in total.

The classifier was then used as an automatic special events detector. An energy-based algorithm was also implemented for comparison purposes. In this algorithm, the energy of the audio signal in each 1s window is calculated and a highlight is detected once the energy level is higher than 40% of the maximum energy level in the video clip. The accuracy of the classifiers was measured using the Recall and the Precision ratios. Recall that:

$$\text{Recall} = \frac{\text{Real events automatically detected by the system}}{\text{Total real events}}$$

and

$$\text{Precision} = \frac{\text{Real events automatically detected by the system}}{\text{Total events detected by the system}}$$

The results shown in Table 1 confirm the supposition that simple energy-based algorithms are less efficient than our PGM-MLP audio classifier for highlights detection. Generally errors in the energy-based approach are due to the time instants when the commentator speaks, and hence increases the energy of the audio signal.

Nevertheless, the high recall ratio is an indicator that the audio stream can provide important cues for tennis videos segmentation.

Algorithm	Recall %	Precision %
Energy-based	88.1 %	77.1 %
PGM-MLP	97.6 %	87.2 %

Table 1 Highlights detection accuracy for Energy-based algorithms and the proposed PGM-MLP approach.

5. SOCCER AUDIO ANALYSIS

Soccer audio analysis consists of detecting special events based on the audio stream only. Several audio classes in soccer audio can be detected including: whistles, speech, and crowd. However, we use the audio classifier to classify soccer audio stream into action or no-action. The action is generally transcribed to the listener by excitation in the commentator's voice and in a crowd's noise. Crowd's noise alone can be confusing since it can be directed to give confidence to one team. Thus, the definition of "action" class is based on both crowd's noise and commentator's excitation.

Two applications can benefit from such an action detection in soccer games: 1- Intelligent soccer video player, that highlights the action and the user can then navigate into a match easily Figure 4. 2- Automatic résumé of soccer videos, where a match of 90 minutes can be resumed in 1 min based on action detection results.

The proposed approaches in the literature to special events' detection in soccer videos are generally based on the energy of the signal [16], [19], [20], [25]. Such techniques cannot detect other classes than "high energy class", and "low energy class". Also such techniques need to define a threshold that can be different for different videos. Our audio classifier based approach is adaptable since the user can easily change the definition of the desired classes. Also, the proposed approach is threshold free and is independent of the loudness (or the volume) of the audio signal making it suitable for real world applications where many sound preprocessing techniques are applied to the audio stream which can change its characteristics. Figure 5 shows an example of the energy of the audio signal in a soccer video. One can see clearly that energy-based approach with a simple threshold is not a good solution for action moment detection.

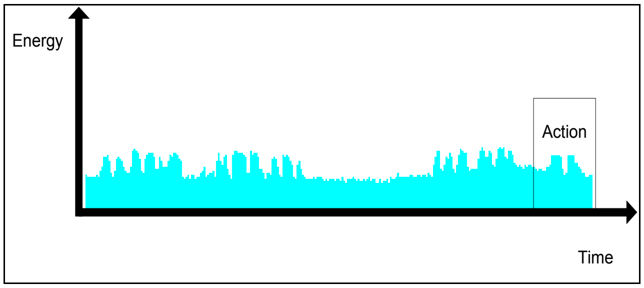

Figure 5 The energy of the audio signal is not always a good indicator of important actions

5.1. Experiments

The database used for the evaluation consists of 3 soccer games from the UEFA. 20 seconds were extracted from a special event in match 1, and 20 seconds from the same match containing normal activity in the audio stream. These 40 seconds constituted the training data for the audio classifier. The audio stream of the 3 matches was classified by the system as action/no-action.

The definition of "action" in soccer match can include subjective judgments making the evaluation using the Recall ratio difficult. Therefore we evaluated the accuracy of the system using the precision ratio. Still, the goals in soccer matches are objectively "action", therefore the recall in goal detection can be easily used in the evaluation.

Since the classifier's output is the probability between 0 and 1 of a frame to be "action", a threshold is needed to decide if a frame will be included in the resume. Clearly the lower the threshold is, the higher the recall ratio and the lower the precision ratio will be. Besides, the duration of the key-moments' duration (or the resume) depends on the threshold. In the experiments the threshold was set to "0.9".

Table 2 shows the duration of the resume, the precision in the resume, and the accuracy of goal detection.

Video	Goals	Goals detected	Precision %	Important Time (s)
Match1	3	3	85 %	90
Match2	0	0	93 %	40
Match3	4	4	88 %	80

Table 2 Goals detection, important time extracted automatically, and the precision of the important time extracted using the proposed audio-based highlights detection approach

The efficiency of the proposed approach for goal detection and the compactness of key moments' duration extracted from soccer videos with relatively high precision are indicators that the audio stream provides valuable semantic information in soccer videos.

6. CONCLUSION

The use of a general audio classifier for highlights detection in sports videos was investigated in this paper. Although our approach has been tested only on two sports types which are soccer and tennis, the application on other types of sports videos is straightforward. Audio-based highlights detection shows its effectiveness for tennis videos, as 97.8% of the end of serves were correctly detected. In the case of soccer videos, goals were detected using audio highlights. The experiments also show that our approach is more efficient than simple energy-based algorithms.

However, the approach proposed in this paper is extremely sensitive to the spectators' and commentators' behaviors. For instance, if the spectators of a tennis match do not appreciate the game our approach will probably fail. One way to overcome such a drawback is to combine audio and image analysis.

7. REFERENCES

[1]. Yong Rui; Anoop Gupta; Alex Acero; *Automatically Extracting Highlights for TV Baseball Program*s, Proc. ACM Multimedia, Los Angeles USA, Pages 105 -115, October 2000

[2]. D. Yow, B.L.Yeo, M. Yeung, and G. Liu, *Analysis and Presentation of Soccer Highlights from Digital Video* Proc. ACCV, 1995, Singapore, Dec. 5-8, 1995

[3]. Xie, L.; Chang, S.; Divakaran, A.; Sun, H., *Structure Analysis of Soccer Video with Hidden Markov Models*, Proc. Of IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP2002 May 2002

[4]. Alejandro Jaimes, Tomio Echigo, Masayoshi Teraguchi, Fumiko Satoh *LEARNING PERSONALIZED VIDEO HIGHLIGHTS FROM DETAILED MPEG-7 METADATA*, Proceedings of IEEE ICIP2002, Rochester, New York, USA, 2002

[5]. Wensheng Zhou, Son Dao, C.-C. Jay Kuo, *On line knowledge and rule-based video classification system for video indexing and dissemination*, Information Systems 27 (2002), pp 559-586, Elsevier Science, 2002

[6]. Y. Gong, T.S. Lim, and H.C. Chua, *Automatic Parsing of TV Soccer Programs*, Proc. Of IEEE International Conference on Multimedia Computing and Systems, pp. 167 – 174, May, 1995

[7]. G. Sudhir, John C. M. Lee, Anil K. Jain, *Automatic Classification of Tennis Video for High-level Content-based Retrieval*, Proc. Of the 1998 International Workshop on Content-Based Access of Image and Video Databases (CAIVD '98), Bombay India, 1998

[8]. G. S. Pingali, Y. Jean and I. Carlbom. *Real-time tracking for Enhanced Tennis Broadcast*s. In the Proceedings of CVPR 98, the IEEE International Conference on Computer Vision and Pattern Recognition, pp. 260-265, June 1998

[9]. Gopal S. Pingali, Agata Opalach, Yves D. Jean, Ingrid B.Carlbom*, Instantly indexed multimedia databases of real world events*, IEEE Transactions on Multimedia, VOL 4, NO 2, June 2002

[10]. Hisashi Miyamori, *Improving accuracy in behaviour identification for content-based retrieval by using audio*

and video information, Proceedings of IEEE ICPR02, VOL 2, pp 826-830, 2002

[11]. J. Assfalg, M. Bertini, A. Del Bimbo, W. Nunziati, P. Pala, *Soccer Highlights Detection and Recognition Using HMMs*, in Proc. of IEEE Int'l Conf. on Multimedia and Expo ICME2002

[12]. Utsumi, Miura, Ide, Sakai, Tanaka, *An object detection method for describing soccer games from video*, Proc. Of IEEE Intl. Conf. on Multimedia and Expo ICME2002, vol.1, pp.45-48,Aug 2002,

[13]. M. Petkovic, W. Jonker, *Content-Based Video Retrieval by Integrating Spatio-Temporal and Stochastic Recognition of Events* , Proc. of IEEE International Workshop on Detection and Recognition of Events in Video, Vancouver, Canada, July 2001.

[14]. Hong Lu, Yap-Peng Tan, *Sports video analysis and structuring*, Proc. Of the IEEE 4th Workshop on Multimedia Signal Processing:, Cannes, France, October 3-5, 2001

[15]. S.-F. Chang, D. Zhong, and R. Kumar, *Real-Time Content-Based Adaptive Streaming of Sports Video*, Proc. of IEEE Workshop on Content-Based Access to Video/Image Library, Hawaii, Dec. 2001

[16]. A. Hanjalic, L.-Q. Xu: *User-oriented Affective Video Analysis*, Proc. Of IEEE Workshop on Content-based Access of Image and Video Libraries, in conjunction with the IEEE CVPR 2001 conference, Kauai, Hawaii (USA) , December 2001

[17]. Z. Liu, J. Huang, and Y. Wang, *Classification of TV Programs Based on Audio Information using Hidden Markov Model*, Proc. Of IEEE Workshop on Multimedia Signal Processing, pp 27-32, Log Angeles, CA, Dec.7-9, 1998.

[18]. J. Kittler, K. Messer, W. Christmas, B Levienaise-Obadia, D. Koubaroulis, *Generation of semantic cues for sports video annotation*, Proceedings of the IEEE ICIP2001 conference, pp 26-29, Thessaloniki, Greece, October, 2001

[19]. S. Dagtas, M. Abdel-Mottaleb, *Extraction of TV highlights using multimedia features*, Proc. Of the IEEE 4th Workshop on Multimedia Signal Processing:, Cannes, France, October 3-5, 2001

[20]. Yuh-Lin Chang; Wenjun Zeng; Kamel, I.; Alonso, R., *Integrated image and speech analysis for content-based video indexing*, Proceedings of the Third IEEE International Conference on Multimedia Computing and Systems, pp 306 -313, 1996

[21]. F.K. Soong, A.E. Rosenberg, *On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition*, IEEE Transactions on ASSP, VOL 36, NO 6, pp 871-879, June 1988

[22]. Simon Haykin, “Neural Networks A Comprehensive Foundation*", Macmillan College Publishing Company,*1994

[23]. Furui S., *Speaker Independent Isolated Words Recognizer Using Dynamic Features of Speech Spectrum*, IEEE Trans Speech Audio Processing. 34, pp 52-59, 1986

[24]. H.Harb, L.Chen, "Segmentation et classification du son", French Patent Pending, BF 02 08 548, July 2002

[25]. R. Leonardi, P. Migliorati and M. Prandini, *A Markov Chain Model for Semantic Indexing of Sport Program Sequences*, Proc of the 4th European Workshop on Image Analysis for Multimedia Interactive Services WIAMIS03, World Scientific Publications, pp 20-27, London, UK 2003