Detection of goal event in soccer videos

Hyoung-Gook Kim, Steffen Roeber, Amjad Samour, Thomas Sikora
Department of Communication Systems, Technical University of Berlin,
Einsteinufer 17, D-10587 Berlin, Germany
Phone: ++49 30 314 28505, fax: ++49 30 314 22514
e-mail: [kim, roeber, samour, sikora]@nue.tu-berlin.de

ABSTRACT

In this paper, we present an automatic extraction of goal events in soccer videos by using audio track features alone without relying on expensive-to-compute video track features. The extracted goal events can be used for high-level indexing and selective browsing of soccer videos. The detection of soccer video highlights using audio contents comprises three steps: 1) extraction of audio features from a video sequence, 2) event candidate detection of highlight events based on the information provided by the feature extraction Methods and the Hidden Markov Model (HMM), 3) goal event selection to finally determine the video intervals to be included in the summary. For this purpose we compared the performance of the well known Mel-scale Frequency Cepstral Coefficients (MFCC) feature extraction method vs. MPEG-7 Audio Spectrum Projection feature (ASP) extraction method based on three different decomposition methods namely Principal Component Analysis(PCA), Independent Component Analysis (ICA) and Non-Negative Matrix Factorization (NMF). To evaluate our system we collected five soccer game videos from various sources. In total we have seven hours of soccer games consisting of eight gigabytes of data. One of five soccer games is used as the training data (e.g., announcers' excited speech, audience ambient speech noise, audience clapping, environmental sounds). Our goal event detection results are encouraging.

Keywords: Goal score Detection, Highlight events Detection, MPEG-7, Mel-scale Frequency Cepstrum Coefficients (MFCC), HMM

1. INTRODUCTION

Research towards the automatic detection and recognition of events in sport videos data has attracted a lot of attention in recent years. Soccer video analysis and events/highlights extraction are probably the most popular topics in this research area. Soccer is a very popular sport, the whole game is quite long, often there are several games being played on the same day or at the same time and viewer may not be able to watch all of them. Users desire the capability to watch the programs time-shifted (on-demand) and/or desire to watch only the highlights such as goal events in order to save time. Recently, audio contents become more and more important clues for detecting events relating to highlights across different sports, because different sounds can indicate different important events. There have been many works on integrating visual and audio information to automatically generate highlights for sport programs. In [1], a shot-based multi-modal multimedia data mining frame work for the detection of soccer goal shots was presented. Multiple cues from different modalities including audio and visual features are fully exploited and used to capture the semantic structure of soccer goal events. In [2], a method to detect and recognize soccer highlights using Hidden Markov Models (HMM) was proposed, in which each model is trained separately for each type of event. The use of HMM classifier can automatically find the temporal change character of the event instead of rule based heuristically modelling which map certain keyword sequence into events. Due to the fact that fast camera movement is associated with soccer highlight points, MPEG motion vectors were used in [3] for highlight detection. A list of video segments is extracted to represent a specific event of interest using the maximum likelihood criterion. Low-level audio descriptors are extracted to order the candidate video segments within the list so that those associated with the event of interest and in order to reduce false alarms.

In this paper we propose a goal event detection method that automatically labels certain segments in a soccer video, which contain highlights, goal events or goal scores. We focus on detecting highlights using audio track features alone

without video track features. The following two reasons give us the motivation. First, visual information processing is computation intensive, and we want to target set-top boxes that have limited available computational power. Second, we want to see, how suitable the use of only audio information is for detecting highlight events in soccer videos. An important role plays the choice of appropriate representation of the audio information and the selection of suitable features that can be extracted from the audio track of the video stream in order to perform a successful detection. An assortment of audio features have been proposed in the literature that can characterize audio signals. To detect main highlight events, only a few of these features are useful. However, the Mel-scale Frequency Cepstral Coefficients (MFCCs) approach is one of the most widely used features for audio classification. Recently, MPEG-7 standard has adopted dimension-reduced, de-correlated spectral features for general sound recognition.

In this paper we compared MFCC and MPEG-7 ASP in order to find the best suitable feature for the overall system and finally we chose MFCC to evaluate our system. We create Hidden Markov Models from the extracted features to detect the event candidate. A combined Hidden Markov Model of an announcers' excited speech and background crowd noise is trained to select goal event segments and to finally add them to the summary. The experiments show that encouraging results can be achieved with audio information only. The extracted goal events can be used for high-level indexing and selective browsing of soccer videos.

This paper is organized as follows. Section 2 gives an overview of the system, describes briefly the applied audio features and introduces the structure analysis mechanism. Experimental results are presented and discussed in section 3. Section 4 provides the concluding notes.

2. ARCHITICTURE OF THE FRAMEWORK

The detection of video event highlights using audio contents comprises three steps: 1) audio feature a video sequence, 2) event candidate detection of main events using HMM, 3) goal event segment selection to finally determine the video intervals to be included in the summary. The architecture of our system is shown in Figure 1. A combined Hidden Markov Model (HMM) of announcers' excited speech and background crowd noise is trained in advance and is later used to find the best time-aligned event candidate sequence.

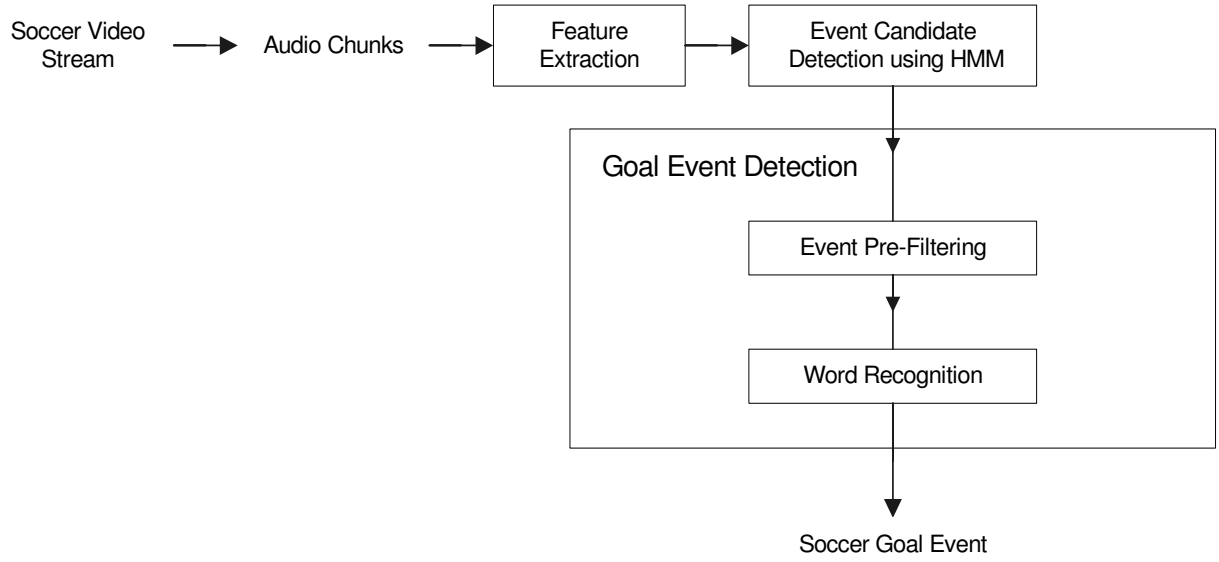

Figure 1: The architecture of detection of goal events in soccer videos.

For the detection of goal events in soccer games we perform first audio segmentation to extract the audio sequences that include candidate events. To do that the audio signals recorded from raw soccer video sequences are divided into sub-segments of 1 second length without overlapping. Each sub-segment is further divided into small overlapped frames

in order to extract vectors of features. Then a HMM is trained from the extracted audio feature vectors. Detection of event candidates is achieved by using the Viterbi algorithm to determine the maximum likelihood state sequence through the HMM. The candidate events contain exciting highlights (such as goal events) and also some non-exciting highlights. To detect the goal events from these candidate events we use a two step procedure. First, we perform an event pre-filtering to remove those candidate event segments which are shorter than a predefined threshold and contain non-exciting highlights. Second, a combined HMM is trained to classify the excited speech of the announcer into two classes such as "goal" or "goal score". Finally, the video intervals of the audio segments that contain true goal highlights are included in the summary.

In order to evaluate the proposed feature sets, left-right hidden Markov Models (HMMs) with seven states are trained using a maximum likelihood estimation procedure known as the Baum-Welch algorithm.

2.1. Feature Extraction

A variety of audio features have been proposed in the literature that can serve the purpose of audio track characterization [4]. 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, brightness and timbre are examples of these features. The physical features such as zero crossing rate, MFCC, energy and spectral centroid are further grouped into spectral features and temporal features according to the domain in which they are calculated [5].

To detect main highlight events, only a few of these features are useful. The Mel-scale Frequency Cepstral Coefficients (MFCCs) approach was for a long time widely used on the area of speech recognition. it also one of the most used features for audio classification. Lately, the Audio Spectrum Projection (ASP) features was standardized in MPEG-7 for general audio classification. Audio Spectrum Projection features are dimension reduced and de-correlated spectral features. Because the feature extraction has a high influence on the recognition rate we evaluate the performance of MPEG-7 audio spectrum projection (ASP) features vs. MFCC according to reduced dimension.

2.2. MFCC Features

The extraction method of MFCC is presented in Figure 2(a). The audio signal is divided into windowed frames for taking the Fast Fourier transform . After a Fast Fourier transform, 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:

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

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.

2.3. MPEG-7 Audio Spectrum Projection (ASP)

Figure 2 (b) shows the extraction method of the MPEG-7 features, the first 2 steps of calculation are the same for both extraction methods. The power spectrum is then transformed to log-scale octave bands to produce (the Audio Spectrum Envelope ASE) a reduced representation of the spectrogram of the original audio signal, the ASE is converted to the decibel scale. Each decibel-scale spectral vector is normalized with the RMS (root mean square) energy envelope, thus yielding a normalized log-power version of the ASE called NASE and represented by $L \times F$ matrix. It is defined as:

$$X(l,f) = \frac{10\log_{10}(ASE(l,f))}{\sqrt{\sum_{f=1}^{F}\{10\log_{10}(ASE(l,f))\}^2}}\,, \tag{2}$$

where $l(1 \le l \le L)$ is the time frame index, $f(1 \le f \le F)$ is the logarithmic frequency range, L is the total number of frames and F is the number of ASE spectral coefficients.

Finally, a basis decomposition method such as PCA, ICA or NMF is applied in order to calculate the MPEG-7 projection features.

For the basis decomposition step of MPEG-7 ASP feature extraction, Principal Component Analysis (PCA), Independent Component Analysis (ICA), or Non-negative Matrix Factorization (NMF) are used in order to reduce the dimension of the feature space while retaining as much important information as possible.

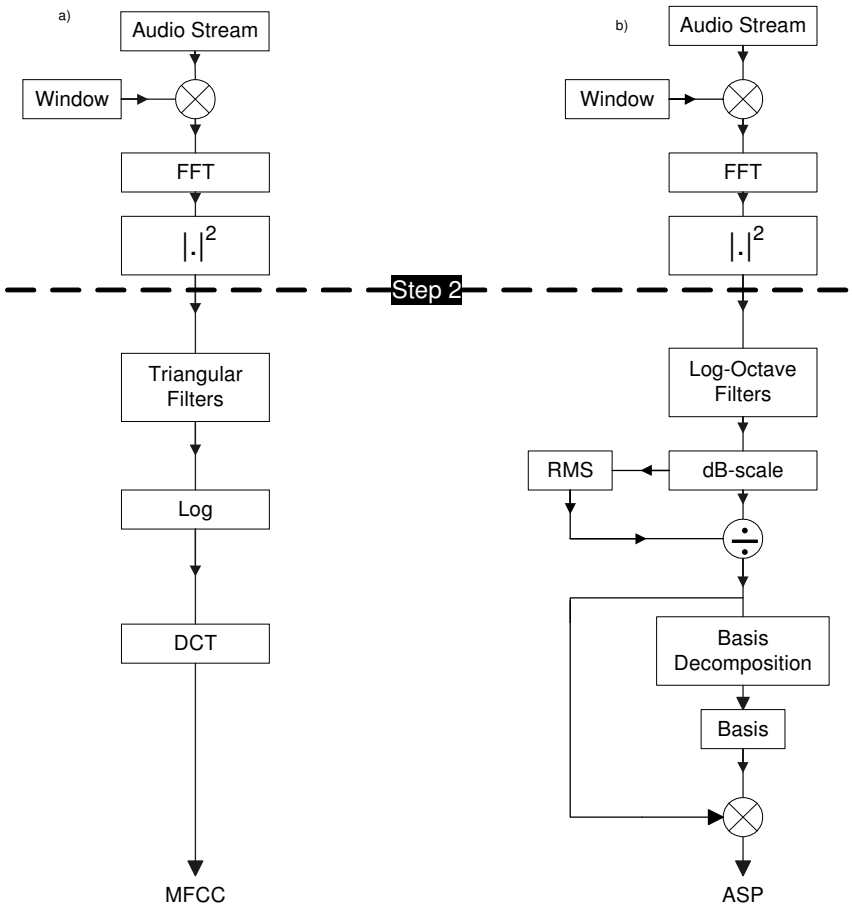

Figure 2: (a) extraction of MFCC (b) extraction of MPEG-7 ASP- features.

- **Independent Component Analysis (ICA)**

Independent component analysis (ICA) is a statistical method for linear transforming an observed multidimensional random vector X into a random vector Y whose components are stochastically as independent from each other as possible [7]. The goal of (ICA) is to find a representation $Y=MX$ in which the transformed Y_i are the least statistically dependent. In ICA, the (pseudo) inverse A of M is called the mixing matrix. ICA de-correlates not only the second order statistics but also reduces higher-order statistical dependencies. Thus, ICA produces mutually uncorrelated basis.

- **Principal Component Analysis (PCA)**

The goal of PCA [6] is to find the best K-dimensional of the data in the least squares sense. PCA de-correlates the second order moments corresponding to low frequency properties and extracts orthogonal principal components of variations.

For a set of N training vectors $X = \{x_1, ...x_N\}$ the mean and covariance matrix can be calculated.

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n \qquad \text{for } n = 1,2,.....N\,, \tag{3}$$

where μ is the mean value of the data vector x_n. The covariance can be calculated as follows:

$$\Sigma = \frac{1}{N}\sum_{n=1}^{N} (x_n - \mu)(x_n - \mu)^T \tag{4}$$

The *K*-dimensional representation of the original signal is given by the projection:

$$y = E^T (x - \mu)\,, \tag{5}$$

where E is the projection matrix composed of the eigenvectors of Σ with the highest eigenvalues.

- **Non-Negative Matrix Factorization (NMF)**

NMF attempts a matrix factorization in which the factors have non-negative elements by performing a simple multiplicative updating. NMF gives a parts based representation of the original dataset [8]. Given an original data matrix as $n \times m$ matrix *V*, where each column contains the n data values for one of the m original data vectors, then the data matrix *V* is approximated by NMF as:

$$V \approx WH = \sum_{a=1}^{r} W_{ia} H_{a\mu} \tag{6}$$

where the rank of factorization, r, is usually chosen as $nm > nr + rm$ to compress the original data into *WH*.
The dimension of the factorized matrices *W* and *H* are $n \times r$ and $r \times n$, respectively. Each column of matrix *W* contains a basis vector and the matrix *H* represents the coefficients for reconstructing the original data matrix *V*.

In last step of the MFCC extraction, a discrete cosine transform (DCT) is applied to the logarithm of the filter bank outputs resulting in vectors of de-correlated MFCC features.

The processing steps of creating MFCC features are compared to the corresponding MPEG-7 ASP extraction steps in Table 1.

	MFCCs	MPEG-7 ASP
1	Convert to Frames	Convert to Frames
2	For each frame, obtain the amplitude spectrum	For each frame, obtain the amplitude spectrum
3	Mel-scaling and smoothing	Log-scale octave bands
4	Take the logarithm for normalization	Normalization with Root Mean Square (RMS) Energy Envelope
5	Take the discrete cosine transform (DCT)	Perform basis decomposition using PCA, ICA, or NMF for projection features

Table 1: Comparison of MPEG-7 ASP and MFCCs.

2.4. Event Candidate Detection Using HMM

Our event candidate detection is based on a model of highlights. In the soccer videos, the sound track mainly includes the foreground commentary and the background crowd noise. Based on the observation and prior knowledge, we assume exciting segments are highly correlated with announcers' excited speech and the audience ambient noise can also be very useful, as audience viscerally react to exciting situations. To detect the goal events we use one acoustic

class model such as announcers' excited speech, audience's applause and cheering for goal or shoot. The HMM is trained with approximately 3 minutes of audio.

Once an ergodic HMM with 7 states is trained by using the well-known Baum-Welch algorithm, it is used for segmenting audio into the event segments by applying the Viterbi algorithm. The Viterbi algorithm determines the most likely sequence of states through the HMM and returns the most likely classification/detection event-label for the sub-segment.

2.5. Event Segment Selection

The candidate events contain not only true exciting highlights but also some non-exciting highlights. However, it is not possible to identify the goal event. In the actual soccer games, goal events never appear in short sub-segments. Instead, they appear in a much longer unit because most exciting segments in soccer occur right after a shoot or a goal. Therefore a pre-filtering process is needed to select a small set of candidate goal sub-segments. Experimentally, we find that a goal segment has a minimum length of 10 seconds.

The event pre-filtered segment (>10 seconds) does not provide enough confidence in extracting the goal event. To detect the goal event we use the sub-system for excited speech classification. The excited speech classification is composed of two steps:

1) noisy environment speech endpoint detection: In TV soccer programs, the noise presence can be as strong as the speech signal itself. To distinguish speech from other audio signals (noise) we use a noise reduction method based on smoothing of the spectral noise floor (SNF) [9]. The noise reduction is very simple but achieves a good tracking capability for a non-stationary noise. The enhanced speech is free of musical tones and reverberation artifacts and sounds very natural compared to methods using other short-time spectrum attenuation techniques.
2) classification excited speech using HMMs: Before the classification we build two models (e.g., excited speech included "goal" and "score" vs. non-excited speech). Starting with these, model-based classification performs a more refined segmentation to detect the goal events.

3. EXPERIMENTS

3.1. Data Set

To validate the effectiveness and robustness of the proposed approach, we collected five soccer game videos from various sources. In total we have seven hours of soccer games consisting of eight giga bytes of data. They come from different sources, digitized at different studios, sampled at 22.05 kHz, and reported by different announcers. One of five soccer games is used as the training data (e.g., announcers' excited speech, audience ambient speech noise, audience clapping, environmental sounds).

3.2. Feature Extraction

The audio signals recorded from raw soccer video sequences are divided into sub-segments of 1 second length without overlapping. To apply the feature extraction, each sub-segment is further divided into 40ms frames with a 50 % overlap between the adjacent frames. Each frame is multiplied by the hamming-window function and transformed into frequency domain using the Fast Fourier Transformation (FFT). For MPEG-7 features the bands are logarithmically distributed between 62.5 Hz lower boundary and 16kHz upper boundary. The lower-upper range has been chosen to be an 8 octave interval, logarithmically centered on 1kHz. The frame is represented by a feature vector. This vector is then projected into the first 7, 13, 23, and 30 components of the PCA, ICA or NMF space of every class. For our experiments Mel-frequency cepstral coefficients are calculated from 17 linear channels under 1kHz and 23 logarithmic channels between 1kHz and 8kHz. In order to compare MPEG-7 ASP- features with MFCC- features, we chose four different dimensions which were 7, 13, 23, and 30 of the feature vector for each of the feature extraction methods.

3.3. Experimental Results

In order to find the best suitable feature for the overall system, we compared the performance of MPEG-7 ASP features and MFCC applied to sport genre audio classification.

We collected 100 audio clips from “Sound Ideas” general sound effects library and TV broadcasting of sport games. Each of them is hand-labeled into one of 5 classes: baseball, basketball, boxing, golf and tennis which is composed of mixed audio with background interference such as applaud and loud cheering. The clip duration was between 2 seconds and more than 10 seconds. The data was divided into training (70 % of data) and testing (30 %) sets. We performed experiments with different feature dimensions 7, 13, 23, and 30 for each of the feature extraction methods. The results of sport genre audio classification are shown in Table 2.

Feature Extraction	Feature Dimension			
	7	13	23	30
ASP onto PCA	87.94 %	89.36 %	84.39 %	83.68 %
ASP onto ICA	85.81 %	88.65 %	85.81 %	63.82 %
ASP onto NMF	63.82 %	70.92 %	80.85 %	68.79 %
MFCCs	82.97 %	88.65 %	93.61 %	93.61 %

Table 2. Comparison of the recognition rates for 4 feature extraction methods.

Results show that in general, it is possible to recognize which one of the 5 sport genres is present on the audio track. With feature dimensions of 23-30 a recognition rate of more than 90 % can be achieved. MFCC features yield better performance compared to MPEG-7 features based on several basis decompositions in the feature dimension 23 and 30.

Table 3 presents the results of speed comparison between MPEG-7 ASP onto several basis decomposition algorithms -such as PCA, ICA, NMF- and MFCC feature extraction. The 23 features for each frame are calculated for the test signal with 4 minute duration.

Feature Dimension 23	Feature extraction Methods			
	PCA	FastICA	NMF	MFCC
ASP onto PCA	75.6s	77.7s	1 h	18.5s

Table 3 Speed comparison between MPEG-7 ASP and MFCC feature extractions.

While the MFCCs of a test example are the same for all audio classes, because there are no DCT bases, the MPEG-7 features of the same example are different. Since each PCA, ICA or NMF space is derived from the training examples of each training class, each class has its distinct PCA, ICA or NMF space. During training, the extraction of the MPEG-7 audio features requires more memory to buffer the features of all the training examples of the audio classes. During testing, the PCA projection needs to be performed for each class. The extraction of MFCC only requires buffering the features of one training example. The features are the same for different classes. For this reason MFCC extraction method is simpler, significant faster and consumes less memory and time than the MPEG-7 ASP feature extraction method.

To perform a NMF, the divergence update algorithm was iterated 200 times. The spectrum basis projection using NMF is very slow in comparison to PCA or Fast ICA. For these reasons we use the MFCC feature extraction method for our detection system of goal events in soccer games.

As shown in Table 4, the goal event detection result is encouraging. Seven out of eight goals from four soccer games were correctly identified, one goal event was misclassified.

Total Goal	8
Identified Goal	7
Miss-identified Goal	1

Table 4 Testing result of detection of goal event.

We developed a highlight and goal events detection tool. A screen shot of our tool is shown in Figure 3. To facilitate user interaction, the left side of the screen includes three interactive buttons: announcer, highlights and goal events. Each button is further integrated into play, next and previous on the right side of the screen.

Figure 3: Screen shot of the proposed automatic extraction of goal events in soccer videos.

4. CONCLUSION

A scheme for automatic detection and recognition of goal events in soccer videos based on audio content analysis was presented in this paper. Our goal event detection result shows that Seven out of eight goals from four soccer games were correctly identified, one goal event was misclassified. The extracted goal events can be used for high-level indexing and selective browsing of soccer videos. A noise reduction method based on smoothing of the spectral noise floor (SNF) was used in order to distinguish speech from background noise. We compared performance and extraction speed of MFCC with performance and extraction speed of MPEG-7 ASP features ASP onto several basis decomposition algorithms such as PCA, ICA, NMF. MFCC features yield better performance than MPEG-7 features based on several basis decompositions in the feature dimension 23 and 30. With feature dimensions of 23-30 a recognition rate of more than 90 % can be achieved.

5. REFERENCES

1. S.-C. Chen, M.-L. Shyu, C. Zhang, L. Luo, M. Chen, “Detection of Soccer Goal Shots Using Joint Multimedia Features and classification Reules”, Proceedings of the Fourth International Workshop on Multimedia Data Mining (MDM/KDD2003), pp. 36-44, August 24-27, 2003, Washington, DC, USA.

2. J. Wang, C. Xu, E. S.Chng and Q. Tian, “Sports Highlight Detection from Keyword Sequences Using HMM”, Proceedings of ICME 2004, June 27-30, 2004, Taipei, Taiwan.
3. R. Leonardi, P. Migliorati, “Semantic indexing of multimedia documents”, IEEE MultiMedia 9 (2) (2002) 44–51.
4. 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, 1998, 20, ½, pp. 61-80, 1998.
5. Vesa Peltonen, Juha Tuomi, Anssi Klapuri, Jyri Huopaniemi, Timo Sorsa “Computational Auditoy Scene Recognition”, ICASSP 2002, Peltonen, V.; Tuomi, J.; Klapuri, A.; Huopaniemi, J.; Sorsa, T.; Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on , Volume: 2 , 2002 Pages:1941 – 1944
6. I. T. Jolliffe, “ Principal component analysis”, Springer- Verlag 1996.
7. U. Amato, A. Antoniadis, G. Gregoire, ”Independent Component Discriminant Analysis”, J. Statistical Society Spain, sottomesso 2000
8. D. Lee and H. Seung. “Learning the parts ofobjects by non-negative matrix factorization”, Nature, 401:788-791, 1999.
9. H.-G. Kim, T. Sikora, “Speech Enhancement based on Smoothing of Spectral Noise Floor” Proceedings of INTERSPEECH 2004 – ICSLP 2004