Audio-Based Event Detection in Videos - a Comprehensive Survey

Rajeswari Natarajan[*1] and Chandrakala.S[#2]
[*1]Department of Electronics and Communication Engineering,
Sri Venkateswara College of Engineering, Irungattukottai, Sriperumbudur, Chennai 602117
rajeswari.natarajan@gmail.com
[#2]Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Thandalam, Chennai 602105
sckala@gmail.com

Abstract—Applications such as video classification, video summarization, video retrieval, highlight extraction, and so forth., need discriminating activities occurring in the video. Such activity or event detection in videos has significant consequences in security, surveillance, entertainment and personal archiving. Typical systems focus on the usage of visual cues. Audio cues, however, contains rich information that might be effectively used for event detection and Multimedia Event Detection (MED) could benefit from the attention of researchers in audio analysis. Many audio-based event detection methods have been proposed for specific applications, while others are generic. This survey presents an exhaustive review of efforts in the past years to address the issues of using audio-based cues in video event detection. Existing methods that are based on audio features and modeling techniques that have been used are summarized in this survey. We hope to provide a good understanding of different directions in this field of research.

Keywords- Event detection, Highlight Extraction, Gaussian Mixture Model, Support Vector Machine

I. INTRODUCTION

In the course of recent years, there has been quick development of manufacturing variety of digital cameras, which has subsequently expanded varying multimedia content. Hence digital and network technologies have yielded an enormous amount of information. Machine learning and understanding of video events is a tough, diverse, interesting and demanding area which has obtained much attention recently. Tasks such as multimedia event detection, scene modeling and discovery of activity patterns have received a lot of attention in the machine learning communities. Variety of applications have been in video surveillance, video summarization, highlight extraction, multimedia semantic annotation, etc.

Such a multimedia content involves multiple information modalities that convey cues related to the properties of underlying events. Various cues are image frames, sound track, text that can be extracted from the video data, spoken utterances that can be decoded from the audio track and so on. Therefore, it is necessary to understand such cues. Until recently, the focus was much on detecting events using visual clues. However, there are various difficulties with visual cues. Visual event may have a huge amount of raw data with richer content which make retrieval or learning task quite difficult. Events have much variety in unrestricted environment. Video quality, presence of noisy data in videos, illumination and feature representation affect the discrimination of the model. Audio cues may contain a wide spectrum of audio events such as vehicle sounds, speech and music. Traditional multimedia information extraction systems have focused mainly on the visual cues such as color histograms, motion vectors and key frames. Audio cues in contrast,can also furnish vivid information which may often represent a theme for video event detection. By only listening to the audio in a video segment, it is usually enough for user to understand what the event is about. They are also useful particularly in circumstances when other modalities neglect to precisely detect events. For instance, distorted visual cues may not reliably help for the task. In such condition, acoustic evidence may complement or enhance the limited visual data available. When both audio and video cues can be reliably extracted and modeled, systems can make detection decisions with higher trust.

Although video based event detection has been more general in literature, few studies on audio-based event detection had been conducted by research community. Video can have different types of aural characteristics such as speech, laughter, applause, cries, environmental sounds, etc. Such mixed audio sound sources has to be dealt with. Gelin and Wellekens [15] addressed video soundtrack indexing using ’phoneme-based keyword spotting’. Saraceno *et al* [52] proposed a method that partitions audio into speech, music, noise and silence. The model performs scene change detection task and classifies video. An innovative method is proposed by Tseng *et al* [61] for semantic video annotation through integrated mining of visual cues, speech cues, and frequent semantic patterns existing in the video. They have shown two main phases 1) Construction of annotation models, such as speech-association, visual-association, visual-sequential and statistical models from annotated videos 2) Fusion of these models for annotating un-annotated videos automatically.

Our main intrigue is in the process of detecting audio-based video excerpts. The problem is quite interesting because it has wide ranging applications in surveillance, sports, entertainment, consumer video and many others. The scope of application is certainly broad. The organization of the paper is as follows: Section II briefly presents the typical types of applications where the task can be carried out. Section III presents general system description. Section IV reviews various audio features used for audio based video event analysis. Section V analyses various modeling paradigms. Future research direction is given Section VI.

II. Types of Applications

Existing research on audio-based video event detection is quite less and yet to be further explored. Audio-based event detection has been executed in diverse fields which are as follows:

A. Surveillance

In surveillance, although visual features may help in detecting events, sounds also may perform better. For example, sounds like gunshots, sudden screams may be required in surveillance. Atrey *et al* [2] has proposed a system for multimedia surveillance using audio features. The approach initially classifies a given audio frame into speech and non speech events. Further normal and excited events are classified using Gaussian Mixture Model (GMM). Four different audio features such as Zero Crossing Rate (ZCR), Linear Prediction Coefficient (LPC), LPC derived Cepstral Coefficient (LPCC), Log Frequency Cepstral Coefficient and then performs further classification into normal and excited events are used. Clavel *et al* [7] have dealt with events detection in noisy environments for a multimedia surveillance. The authors use a class of sounds produced by gun shots. They have aimed at the robustness of the detection and the reduction in false detection.

B. Meeting

Audio features like applause, cheers may be used to retrieve feedback from videos extracted from meeting. Dong *et al* [11] has proposed the automatic recognition of social functional roles in small-group meetings, focusing on a) the significance non-linguistic behaviors, b) the relative time-consistency of the social roles enacted by a given person during the hours of a meeting and c) the happenings and mutual constraints among the roles played by the different people in a social encounter. Comparison of model performance between Support Vector Machine (SVM) and Hidden Markov Model (HMM) has been done. Sidiropoulos *et al* [54] have investigated the problem of segmenting a video into scenes automatically. The approach use high-level audio information, in the form of audio events, for the enhancing the performance of scene segmentation. Also the process has also used the construction of multiple Scene Transition Graphs (STGs) that exploits information coming from different modalities.

C. Sports

Extracting highlights in sports is an interesting application which not only needs visual features but also essentially needs audio cues. A novel framework has been described by Huang *et al* [23] for inferring the low-level structure of a sports game (tennis) using audio track of a video recording of the game. Gaussian Mixture Model and a Hierarchical language model has been used to detect sequences of audio events. A maximum entropy Markov model to used to infer ”match” events from these audio events and multi-grams to understand the segmentation of a sequence of match events into sequences of points in a tennis game. Zhu Liu [37] *et al* has used Hidden Markov Model to classify TV broadcast video. They have used TV programs such as basketball videos, commercials, news , football shows and weather reports for discrimination. Eight frame-based audio features were used to represent the low level audio characteristics and fourteen clip-based audio features were extracted based on these frame-based features to represent the high-level audio properties. An ergodic HMM is built for each of TV programs. The maximum likelihood method is then used for test data to be classified using the models. Xu *et al* [66] presented novel framework that uses audio keywords to assist event detection in soccer video. Audio keywords have been generated from low-level audio features by using support vector machines. The generated audio keywords were used to detect semantic events in soccer video by applying a heuristic mapping. In Xiong and Wang *et al* [63], [64] audio keywords were analyzed using Hidden Markov classifier in order to extract sports (soccer) highlight. [70] detected ball hits in table tennis games using MFCC refined features. The classifier that was used SVM. The results have been compared with data represented which were energy features and their proposed Mel Frequency Cepstral Coefficient (MFCC) refined features.

D. Entertainment

Penet *et al* [48] have explored the use of audio words representations to detect particular audio events such as gunshots and explosions, in order to get more robustness for variations in different audio tracks that are present in Hollywood movies. Each stationary audio segment has been described by one or more audio words obtained by performing product quantization to standard characteristics. In view of these, Bayesian networks are used to capture the contextual information to find audio events in movies. Giannakopoulos [17] handled violence content classification using audio features. Frame level audio features both in time and frequency domain was employed as input to Support Vector Machine. SVM is used for segmenting the violent content. Taskiran *et al* [57] proposed method to summarise videos automatically using transcripts obtained by automatic speech recognition. The full program is divided into segments based on pause detection, segment score derived and on the frequencies of the words and bi-grams it contains. Then, a summary generated are the segments based on the ratio of segments with highest score to the duration. [39] proposed audio features that could suit the scene determination task. They have proved that features influence the result more than clustering method.

E. Consumer Video

Lee *et al* [35] have proposed a strategy for audio-based semantic classification for consumer video. Each video clip is represented as a sequence of MFCC frames. Three clip-level representations such as single Gaussian modeling, Gaussian mixture modeling, and probabilistic latent semantic analysis of a Gaussian component histogram were experimented. Using such summary features, Support Vector Machine classifiers based on the Kullback Leibler, Bhattacharyya or Mahalanobis distance measures are used for classification. Muneesawang *et al* [43] have proposed content-based video retrieval using the combination of audio and visual cues. Adaptive video indexing technique is used to extract the visual feature that emphasizes spatio-temporal information within video clips. A statistical time-frequency analysis that transforms Laplacian mixture models into wavelet coefficients is used to extract audio features.

F. Annual Evaluation of TRECVID data set

Jin *et al* [28] has shown several frameworks for accomplishing MED using only audio data. Multimedia Event Detection is an annual task in the NIST TRECVID assessment. Participants build indexing and retrieval systems for identifying videos in which few predefined events are shown.

G. Other Applications

Nahijima *et al* [46] presented a quick and precise Motion Pictures Experts Group (MPEG) audio classification algorithm based on sub band data domain. Classification task was carried out for 4 segments such as silent, music, speech and applause segments for 1 second unit. Later Bayesian discrimination method for multivariate Gaussian distribution was used for classification task.

III. SYSTEM DESCRIPTION

Digital video is a one that is generated from the camera which is in the form of pixels. A digital video is a sequence of images, called frames displayed at a frame rate, to create an illusion of animation. Frame rate can be defined as number of unique consecutive frames produced per second. Frame rate varies between several standards. A typical video has a frame rate of 25 frames per second. A complete video is partitioned into acts. Each act is further partitioned into scenes. A scene is a sequence of actions where each consecutive frame differs with slight change. Audio is now extracted from the given video either at short-term frame level or at long-term clip level. Data representation of extracted audio signal addresses the issues of representing the examples to be classified in terms of feature vectors. The intention of modeling is to find a mapping from the feature space to the target labels so as to reduce the prediction error. The general system components of the audio based video event detection is presented in Figure 1. The major components of the system are the audio data representation and learning methodologies. An audio signal can be represented by many number of features. Audio feature extraction is an important phase for any type of audio data. It is process of refining enormous amount raw audio data into compact representations which holds the higher level information of the audio. There exists large number of audio features, that are suitable for various such as audio retrieval, audio segmentation, music and information retrieval, environmental sound retrieval, etc. Eyben *et al* [13] presented various feature extractor tools which can be used for many applications. Some of the tools to extract audio features are shown in Table 1. Modeling multimedia data can be organized based on the desired output of the algorithm or on the type of input examples. Supervised learning is a method of learning where the algorithm is trained on the labeled examples. The classifiers such as Gaussian Mixture Model, Support Vector Machine fall under this category. Unsupervised learning is a method of learning where the labels of the input examples are not known a priori.

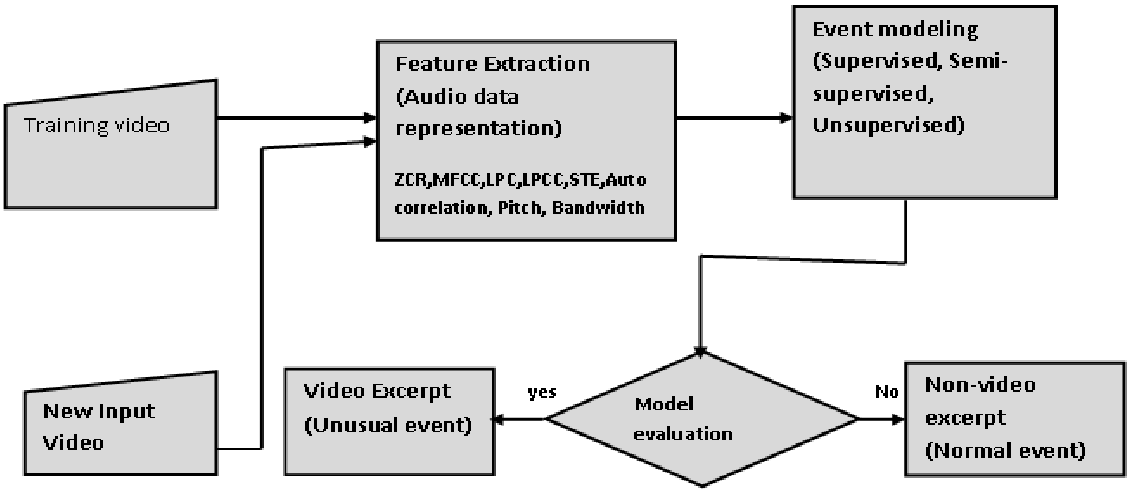
Training video
New Input Video
Feature Extraction (Audio data representation)
ZCR,MFCC,LPC,LPCC,STE,Auto correlation, Pitch, Bandwidth
Event modeling (Supervised, Semi-supervised, Unsupervised)
Model evaluation
yes
No
Video Excerpt (Unusual event)
Non-video excerpt (Normal event)

Figure. 1: General Framework of Audio-Based Video Event Detection System

IV. AUDIO FEATURE ANALYSIS

Audio features depict specific attributes of audio signals. Earlier concept of audio frames were used in speech processing, where analysis varies over a short duration of time typically in milliseconds. For semantic meaning to be explored, analysis has to done for a long duration of time, typically in seconds. Thus audio features can be extracted in two levels: frame level for short duration and clip level for long duration.Several such audio features have been presented in the literature for learning task. Different versions exist for the categorization of audio features. A list of audio features, types, references and applications is listed in Table II. They can be classified into temporal, physical, perceptual, cepstral, modulation features which are briefly explained in following subsections.

TABLE I
Few Audio Feature Extraction Tool-kits

Tools	References
jAudio	[24]
Opensmile	[47]
PRAAT	[49]
YAAFE	[68]

A. Temporal Features

The temporal domain is the native domain for audio signals. Temporal features are the features that are directly extracted from raw audio signal without any preceding information. The processing time for temporal features is slower. Zero Crossing Rate (ZCR), Amplitude based features and power based features fall under the category of temporal features.

TABLE II
Audio Features used in Audio-Based Event Detection in Videos

References	Domain	Features	Applications
[2],[4],[26],[36],[37],[48], [51],[52],[63] [66],[69],[70]	Temporal domain	Zero Crossing Rate, Amplitude, Power	Speech/music discrimination,musical genre classification, highlight detection, singer detection,Environmental sound recognition,Recognition of animal sounds
[2],[26],[39],[58], [63],[66],[69]	Frequency domain(physical)	Auto regression,Adaptive Time Frequency decomposition, Short time Fourier transform	Automatic speech recognition, Audio segmentation, Speech/Music discrimination, Musical genre classification, Music information retrieval, Musical instrument recognition, Language recognition
[4],[23],[39],[37], [51],[59],[69]	Frequency domain(Perceptual)	Brightness, Tonality, Loudness, Pitch, Chroma, Harmonicity	Audio similarity analysis, Rhythm tracking, Music Information Retrieval, Speech/Music segmentation, Audio fingerprinting, Speech/Silence detection, Musical instrument discrimination
[19],[25],[28],[29],[35], [48],[70],[51],[56],[63],[66]	Cepstral domain	Perceptual filter bank, Advanced auditory model, Auto regression	Speech, music and Environmental sound analysis
[59]	Modulation Frequency	Rhythm	Speech/Song discrimination

1) Zero Crossing Features:

Zero crossings are a basic property of an audio signal that is mostly used in learning tasks. Zero Crossing Rate is defined as the number of zero crossings in the temporal domain within a second. Kedem [30] defined ZCR as the measure of dominant frequency in the signal. ZCR is the common feature that is used for music/speech discrimination due to its simplicity. It is also used in other audio domains such as highlight detection [6], speech analysis [8], singer [69] and environmental sound detection [5]. Linear Prediction Zero Crossing Ratio (LP-ZCR) is defined as the ratio between the zero crossing count of the waveform and the zero crossing count of the linear prediction analysis filter [12]. These features help to discriminate between speech and non-speech audio signal. Zero Crossing Peak Amplitudes (ZCPA) has been presented by Kim *et al* [31] ,[32] which is highly suitable for speech recognition in noisy environments. It is an approximation of the spectrum which is directly computed from the signal. Pitch Synchronous Zero Crossing Peak Amplitudes (PS-ZCPA) is an extension of ZCPA which also considers pitch information. It is found to be more robust to noise than ZCPA.

2) Amplitude Features:

Features that are computed from the amplitude of the signal directly are easy and its computation time is fast. MPEG-7 audio waveform is a descriptor that better describes the shape of a waveform on computing the maximum and minimum samples within non-overlapping frames. They are suitable for comparing waveform. Amplitude Descriptor (AD) are the ones that have been developed to recognize sounds of the animal. The descriptor is the one that separates the signal with low and high amplitude by an adaptive threshold. AD signifies the waveform in both quiet and loud segments.

3) Power Features

The energy of a signal is defined as the square of the wave form’s amplitude. The power of a sound is defined as the energy transmitted per unit time or it is the mean-square of a signal. Short Time Energy (STE) is mainly used in fields of retrieving audio. Volume is another important feature which is used to detect silence and also used in segmenting music or speech. Volume is the root mean square of magnitude of the signal within a frame.

B. Physical Features

The most common methods that are used to represent audio features in frequency domain are Fourier transforms and auto correlation. Other methods like Cosine transform, Wavelet transform and Q-transform are also used. Frequency features can be divided in two sets such as physical features and perceptual features.

1) Auto-Regression Based Features:

In auto regression analysis a linear predictor finds the value of each sample which is represented by a linear combination of previous values. Linear Predictive Coding considers source filter model of speech production. It estimates the basic parameters of speech signal which are formant frequencies and vocal tract function. LPC is used for automatic speech recognition, audio retrieval and audio segmentation. Line spectral frequencies (line spectral pairs) are estimated by breaking down the linear prediction polynomial into two separate polynomials. The line spectral frequencies are at the roots of resultant two polynomials. They are extensively used in machine learning applications, recognition of instruments, and speaker segmentation task.

2) Adaptive Time Frequency Decomposition Based Features:

Features of this category are based on the transform coefficients. Adaptive time frequency transform features were proposed by Umapathy *et al* [62]. The signal is broken down into a set of Gaussian basis functions of many scales, translations and center frequencies. These features are mainly used in discriminating music genres.

3) Short Time Fourier Transform Based Features:

Short Time Fourier Transform contains real and complex fields. The real fields represent distribution of the frequency components while the complex fields contains information on the phase of components. The features that possess the properties of spectral envelope are 1) Sub band energy ratio that are used in music analysis and audio segmentation, 2) Spectral flux mostly used in music/speech discrimination, music information retrieval, speech analysis, 3) Spectral slope that are used to discriminate speech and non-speech segments, 4) Spectral peaks that are used to identify a short segment of music, 5) Group delay function which are used in music analysis, 6) Modified group delay function which are used in recognition of languages [44] [45], phonemes and syllable detection.

C. Perceptual Features.

Features that have a semantic meaning in the context of human auditory perception are called Perceptual Frequency Features. Brightness, Tonality, Loudness, Pitch, Harmonicity are commonly used perceptual frequency features. A signal is composed of both low and high frequencies. A sound becomes brighter due its high frequency content and silence is less dominant due its less frequency content. Tonality is property of sound that discriminates tonal sounds from noisy sounds. Tonality measures can be categorized into bandwidth measures and flatness measures. Loudness features are the ones that signifies the auditory sensation. Loudness measure can be used for audio retrieval. Pitch is dimension of sound which can be loudness, duration and timbre. Pitch is commonly related to chroma and harmonicity. Chroma is the one that is divided into 12 pitch classes, where each corresponds to one note of 12-tone equal temperament. Harmonicity is a characteristic that differentiates harmonic sounds from in-harmonic sounds.

D. Cepstral Features.

Cepstral frequency are log magnitude representations in which the frequency are smoothed and they possess timbral properties and pitch. They have orthogonal basis which helps in performing similarity comparisons. These are widely used all of audio extraction. The most popular cepstral features used are Mel-Frequency Cepstral Coefficient, Δ MFCC(first order derivative), $\Delta\Delta$ MFCC second order derivative, Bark Frequency Cepstral Coefficient (BFCC), Homomorphic Cepstral coefficient (HCC). They represent timbral properties of a signal. Finding MFCCs involves a conversion of Fourier coefficients to Mel-scale. Later, the resultant vectors are logarithmized and decorrelated by Discrete Cosine Transform (DCT), which helps in removing redundant information.

V. MODELLING PARADIGMS

An event can be defined as any human-visible occurrence that has importance to represent video contents fused with audio. Each video can consist of many events. Current research aims at models that handles this problem. Classification is a technique of modeling a set of labeled instances(training) and then to classify a test instance into one of the classes using model. Table III shows various model paradigms used in literature and are explained in following sub-section.

A. Hidden Markov Models.

Hidden Markov models have been extensively used for modeling the temporal dynamics of varying length patterns of short duration. A HMM is a finite state machine characterized by the number of states in the model, the state-transition probability distribution, the observation symbol probability distribution for each state, and the initial state probability distribution. Continuous density HMMs use probability densities to represent the continuous observation distributions of the states. The continuous observation density for a state is estimated by assuming that it can be represented by a mixture of Gaussian density functions. Then the estimation of continuous density for a state involves estimation of the mean vector and co variance matrix of each component of the Gaussian mixture and the estimation of the mixture coefficients. The HMM for a class is trained using the varying length sequences corresponding to the sequences of feature vectors of multiple examples of the class. The HMM for a class is trained to maximize the likelihood of the model generating the sequences of that class. During recognition, the sequence of a test pattern is given as input to the HMM of each class, to compute the probability of the test sequence being generated by that model. Then the class of the model with the highest probability is assigned to the test pattern. The Hidden Markov Model [3] describes a Markov chain on latent variable $h_{1..T}$ The visible variables are dependent on the latent variables through $p(o_i/h_i)$. The joint distribution is defined by

$$p(h_{1..T}, o_{1..T}) = p(o_1/h_1)\prod_{i=2}^{T} p(o_i/h_i)p(h_i/h_{i-1}) \tag{1}$$

For a stationary HMM the transition $p(h_i/h_{i-1})$ and $p(o_i/h_i)$ are constant over time. HMM has good ability to capture the temporal statistical property of stochastic process and is used widely in machine learning field. Liu *et al* [38] use HMM to classify broadcast news by using eight frame based audio features depicting low level audio cues and fourteen clip level audio features depicting high level audio features. Wang *et al* [63] detect highlights from keyword sequences using HMMs. Xiong *et al* [65] create audio labels using Gaussian Mixture Model and and video labels are obtained using quantization of the mean motion vector magnitudes initially. Later highlights of sports using discrete-observations Coupled Hidden Markov Model(CHMMs) on video and audio labels are classified using a enormous training set of broadcast sports highlights. HMM has the ability to exhibit some degree of in-variance to local warping of time axis. HMM can be trained effectively using maximum likelihood when training sequence is sufficiently long. Assuming that distribution of individual observation parameters well represented as a mixture of Gaussian or auto regressive densities is a limitation. Probability of being in a given state at time '*t*' only depends on the state at time '*t-1*' is inappropriate, where dependencies extends through many states. Choice of type of model (ergodic or left-to-right), choice of model size (number of states), choice of observation symbols (discrete or continuous) are some issues during implementation.

TABLE III
Modeling Techniques in Audio-Based Video Event Detection

References	**Modeling Framework, Learning Algorithm Techniques**	**Applications**
[2],[23],[26],[35],[29], [36],[56],[18],[53]	Gaussian mixture model	Surveillance,Semantic concept classification in consumer video,Video indexing
[19]	Piece wise Gaussian mixture model	Highlight extraction of games
[26]	K-Nearest Neighbour	Scene detection
[22],[69]	Hidden Markov Model	Audio-Visual event recognition in videos ,Video classification, Sports highlight extraction, Person detection
[37],[58],[63]	Ergodic Hidden Markov model	Video classification
[4],[51],[56],[59], [66],[70]	Support Vector Machine	Semantic video search, Highlight extraction for games, Creation of audio words,Ball hit detection
[25],[69]	Bayesian Network	Violent scene detection
[16]	Variation of Bayesian network	Detecting violent scenes in movies
[43]	Laplacian mixture model	Video retrieval
[22]	Graphical model	Audio-Visual event recognition in videos
[59]	Multilayer perceptron	Semantic video search
[35]	Probabilistic Latent Semantic Analysis	Semantic concept classification
[59]	Hierarchical clustering	Video search
[25],[69],[39]	Clustering	Video search, Video classification and segmentation, Scene Determination

B. *Gaussian Mixture Model.*

The GMM can be considered as an extension of the Vector Quantization (VQ) model, in which the clusters are overlapping. That is, a feature vector is not assigned to the nearest cluster, but it has a non-zero probability of being generated from each cluster. Gaussian mixture model is a linear combination of Gaussian components. Gaussians are especially convenient continuous mixture components as they constitute 'bumps' of probability mass, helping an intuitive elucidation of the model. Assume the data distribution is Gaussian. For a 'd' - dimensional feature vector **x**, the likelihood of **x** for a GMM λ with *K* components is defined as follows

$$p(\boldsymbol{x}) = \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{w_k}\, \mathbf{N}(\boldsymbol{x}/\boldsymbol{\mu}_k\ , \mathbf{C_k}\ ) \tag{2}$$

The components weights,w_k satisfy the constraints, $0 \leq w_k \leq 1$ and $\sum_{k=1}^{K} w_k = 1$. Each of the 'K' uni-modal Gaussian distributions is parametrized by a d - dimensional mean vector, μ_k and a co-variance matrix, C_k as follows

$$N(\boldsymbol{x}/\boldsymbol{\mu_k}, \boldsymbol{C_k}) = \frac{\mathbf{1}}{(\mathbf{2}\boldsymbol{\pi})^{\boldsymbol{d}/\mathbf{2}} \quad |\boldsymbol{C_k}|^{\mathbf{1}/\mathbf{2}}} \boldsymbol{e}^{\frac{-\mathbf{1}}{\mathbf{2}}(\boldsymbol{x}-\boldsymbol{\mu_k})^{\boldsymbol{t}} \boldsymbol{C_k}^{-\mathbf{1}}(\boldsymbol{x}-\boldsymbol{\mu_k})} \tag{3}$$

The parameters of GMM are as follows:

$$\lambda = \{w_k, \mu_{k,} C_k\}, k = 1,2, \ldots \; K \tag{4}$$

Let a multivariate varying length pattern be denoted by a set of feature vectors $\boldsymbol{X}$ = { x_1, x_2, . . ., x_j, . . x_n} where x_j is a d-dimensional feature vector and n is the number of feature vectors. For a varying length patten, $\boldsymbol{X}$, the likelihood score using a GMM with λ as the set of model parameters is defined as follows:

$$p(\boldsymbol{X}/\lambda) = \prod_{j=1}^{n} p(\boldsymbol{x}_j/\lambda) \tag{5}$$

The log-likelihood is given by

$$\ln p(\boldsymbol{X}/\lambda) = \sum_{j=1}^{n} p(\boldsymbol{x}_j/\lambda) \tag{6}$$

Tasks [10] where different examples of same class have different number of acoustic events can be modeled using GMMs. One of the variant of GMM is adapted GMM that be modeled to handle audio data variability in videos. Another variant of GMM is GMM-Universal Background Model used to represent general independent feature characteristics to be compared against a model of example-specific feature characteristics when making an accept or reject decision. This can be used in audio based video verification tasks. Huang *et al* [23] proposed techniques that consist of GMMs and a hierarchical language model to detect sequences of audio events. A maximum entropy Markov model had been used to grasp "match" events from these audio events and multi grams to capture the segmentation of a sequence of match events into sequences of points in a tennis game. Guo *et al* [18] classified internet videos using audio. Three modeling approaches are investigated such single Gaussian, Gaussian Mixture Model with Bag of Audio Words and Support Vector Machine accompanied with different distance measure. GMM representation with Bhattacharya distance has proven to give good results.

GMM is the fastest algorithm for learning mixture models. Estimating the parameters of a full co-variance GMM requires more training and computationally expensive.

TABLE IV
Data Set and Applications in Audio-Based Video Event Detection

References	Dataset	Audio-Based Application in Videos
[2]	Office Corridor video (Talk,Shout,Knock,Footstep)	Video Event Detection
[23]	Wimbledon2008 Tennis	Inferring Structure of the game detection
[69]	Tennis games (Ball hits, background speech, cheering, content of sports games)	Ball hit detection
[26]	TV news, China Central TV (sports and news)	Video Segmentation
[35]	Youtube 1873 videos(25 concepts)	Semantic concept classification
[36]	3 movies	Content-based movie analysis
[19]	Australian open 2002 Tennis (Applause, speech or silence)	Highlight Extraction
[22]	Meeting videos, Surveillance	Audio-Visual event recognition
[37]	TV programs (Commercial, Basketball, Football games, news, weather forecast)	Classification of TV programs
[58]	64 Video clips (falling, walking , walking +talking)	Falling person detection
[63]	Broadcast soccer video FIFA world cup 2002 (Goal, Corner kick, Shot and Goal kick)	Sports highlight detection
[51]	7 hours video	Sports highlight extraction
[66]	European project VIDI-VIDEO	Creating audio keywords for event detection
[4]	FIFA world cup 2002 (Whistler, Commentator, Speech and Audience sound)	Detecting audio events
[16]	Movies	Violent scene detection
[18]	Internet Videos	Video classification
[25]	Kodak Video Benchmark data set, (Birthday, Wedding, Show, Parade)	Video summarization
[43]	Hollywood movies, 15 music recording	Fusion of features gives good result
[39]	German TV movies(Groundhog day, Forest gump)	Video indexing and retrieval
[60]	TRECVID	Annual Evaluation of video data set, Video indexing, retrieval, summarization
[41]	MediaEval(25 genres)	Violent scene detection
[21]	Hollywood movies	Video summarization, retrieval, indexing,annotation
[19]	Australian open tennis	Highlight Extraction
[63]	FIFA worldcup	Highlight Extraction,Video Summarization
[34]	KTH (25 subjects, 6 types of actions, 4 scenarios	Any of video processing task
[33]	Kodak	Consumer video processing task
[9]	CCV (9317 videos, 20 semantics)	Action Recognition

C. Support Vector Machine.

The support vector machine is a kernel based discriminative classifier that focuses on modeling the decision boundaries between classes. The SVM based classifier gives a good generalization performance to classify the unseen data. SVM involves training the examples based on the minimization error function

$$\frac{1}{2}w^t w + C \sum_{i=1}^{N} \xi_i \tag{7}$$

subject to the constraints

$$y_i(w^t \varphi(x_i) + b) \geq 1 - \xi_{i,}, \xi_i \geq 0, i = 1, 2 \ldots N \tag{8}$$

where 'b' is the bias, 'C' is the trade-off parameter, 'w' is the weight, 'φ' is the kernel function that transforms the training examples from the input space to the feature space, 'ξ_i' tells the parameters handling non separable input examples, $\boldsymbol{x}_i$ represents the training examples, $y_i \, \varepsilon \, \{+1, -1\}$ depicts the class label of the i^{th} training example. SVM has always been demonstrated to give exceptional outcomes. Harb *et al* [19] use GMM for extracting sports highlight using audio features. The audio features have been used to build piece wise GMM and Neural Network model. In Lee [35] three clip level representation of MFCC audio features were used to model single-Gaussian Mixture Modeling, Gaussian Mixture Modeling, and probabilistic latent semantic analysis of a Gaussian Component Histogram. Using the measures obtained, Support Vector Machine with different distance measures had been used for classification. Jin *et al* [28] use GMM and SVM to detect event based MFCC audio features. This task is annually performed in order to evaluate TRECVID database.

SVM gives a good generalization even for the unseen data. SVM provides a unique solution, since the optimality problem is convex. By the use of kernel, SVM performs good not only for linearly separable data, but also for non-linear data. Important issues lie in choice of kernel, choice of width in Gaussian kernel. Usage of memory is high as it has to deal with quadratic programming for large-scale tasks.

D. Bayesian Networks.

A Bayesian network is a graphical representation that lets us to depict and reason about an inconstant domain. In Bayesian network, a model is represented using a set of random variables and their conditional dependencies using a directed acyclic graph (DAG). The vertices in Bayesian network depict a set of random variables, $Y = \{Y_1, Y_2, Y_3 \, \dots, Y_i, .., Y_N\}$ from the domain. A set of directed links connect a pair of vertices $Y_i \longrightarrow Y_j$ depicting direct dependencies between the random variables. Fewer the links in the graph, stronger conditional independence properties are present in the network and the model obtained has less degrees of freedom. In Penet *et al* [48] Bayesian networks are used to grasp the contextual information in order to find audio events.

Bayesian network depicts the probabilistic relationships among features. If the dependencies in the joint distribution are sparse, the networks helps in saving the space. When the network is unknown, exploring it is computationally a difficult task.

VI. FUTURE DIRECTION

This review has highlighted the research in audio-based video excerpt detection. Table 2 shows the applications, data sets and references. Representation of the features (audio data extracted from video) is a challenging task. Audio extracted from video is usually a varying length long duration pattern. To measure the similarity or dissimilarity between those patterns is quite non-trivial. Audio-specific kernel can be explored to model the aural characteristics that is extracted from video. In order to establish links between classification of genre, event and object in video classification task, hierarchical semantic relationship between scenes, shots, and key frames in a video can be built utilizing sound characteristics. As audio features are less expensive when compared to visual features, they may be fused with visual features in video segmentation task. In affective based video event detection, audio track can be combined with visual semantics to understand emotional semantics in videos.

VII. CONCLUSION

Video event detection is a main research area in computer vision. But, it has also fascinated analysts in audio field to handle the issue of detecting events using audio cues. In this survey, we have presented different ways in which audio based event detection has been proposed in the literature. The research attention has to involve in coming up with new frameworks that will efficiently perform with real-time environment with so many video events. There are many different directions to handle this problem. Collection of varied techniques used in literature may give rise to enormous applicability in various domains. This might also provide a scope for new directions in this area.

REFERENCES

[1] P. K. Atrey, M. S. Kankanhalli, and R. Jain, ”Information Assimilation Framework for Event Detection in Multimedia Surveillance Systems”,Multimedia Systems, pp239-253, 2006

[2] P. K. Atrey, M. C. Maddage, and M. S. Kankanhalli. ”Audio based event detection for multimedia surveillance.” In Acoustics, Speech and Signal Processing, ICASSP Proceedings, 2006, vol.5, pp. V-V

[3] D. Barber, Bayesian reasoning and machine learning. Cambridge University Press, 2012.

[4] M. Bugalho, J. Portelo, I. Trancoso, T. Pellegrini, and A. Abad. ”Detecting audio events for semantic video search.” In Interspeech, 2009, pp. 1151-1154.

[5] R. Cai, L. Lu, A. Hanjalic, and H. J. Zhang , ”A Flexible framework for audio effects detection and auditory context inference”,IEEE Transactions on Speech and Audio Processing 14 pp. 1026-1039, May 2006

[6] C.C. Cheng and C. T. Hsu, ”Fusion of audio and motion information on HMM based highlight extraction for baseball games” IEEE Transactions on multimedia vol. 8 no.3, pp. 585-599, June 2006.

[7] C. Clavel, T. Ehrette, and G. Richard. ”Events detection for an audiobased surveillance system.” IEEE International Conference on In Multimedia and Expo, 2005, pp. 1306-1309.

[8] Z. J Chuang and C. H. Wu, ”Emotion Recognition using acoustic features and textual content”, IEEE International Conference on Multimedia and Expo, June 2004, Volume 1, pp. 53-56.

[9] Columbia Consumer Video Dataset [Online]. Available: http://www.ee.columbia.edu/ln/dvmm/CCV/

[10] S. Chandrakala, Similarty paradigm based approaches for classification of long duration for classification of long duration speech and audio data in a discriminative framework, Indian Institute of Technology Madras, Chennai, India, 2011.(Ph.D Thesis)

[11] W. Dong, R. Lepri, F. Pianesi and A. Pentland, ”Modelling Functional Roles Dynamics in Small Group Interactions.” IEEE transactions on multimedia vol.15 no.1 pp. 83-95 , 2013

[12] K. El-Maleh, K. M. Petrucci and P. Kabal, ”Speech/Music discrimination for multimedia applications”, IEEE International Conference on Acoustics, Speech and Signal Processing, June 2000, Volume 6, pp.2445-2448.

[13] F. Eyben, M. Wllmer, and B. Schuller, ”Opensmile: the munich versatile and fast open-source audio feature extractor.” ACM ,In Proceedings of the international conference on Multimedia, 2010, pp. 1459-1462.

[14] D. Fuentes, R. Bardeli, J. A. Ortega, and L. Gonzalez-Abril. ”A similarity measure between videos using alignment, graphical and speech features.” Expert Systems with Applications vol. 39, no. 11, pp. 10278-10282, 2012.

[15] P. Gelin, and C.J. Wellekens. ”Keyword spotting enhancement for video soundtrack indexing.” ICSLP, IEEE Fourth International Conference on In Spoken Language, 1996, vol. 2, pp. 586-589.

[16] T. Giannakopoulos, A. Makris, D. Kosmopoulos, S. Perantonis, and S. Theodoridis, ”Audio-visual fusion for detecting violent scenes in videos.” In Artificial Intelligence: Theories, Models and Applications, Springer Berlin Heidelberg, 2010.

[17] T. Giannakopoulos, D. Kosmopoulos, A. Aristidou, and S. Theodoridis. ”Violence content classification using audio features.” In Advances in Artificial Intelligence, Springer Berlin Heidelberg, 2006.

[18] J. Guo, and C. Gurrin. ”Short user-generated videos classification using accompanied audio categories.” ACM international workshop on Audio and multimedia methods for large-scale video analysis, 2012, pp. 15-20.

[19] H. Harb, and L.Chen. ”Highlights detection in sports videos based on audio analysis.” Third International Workshop on Content-Based Multimedia Indexing CBMI September 2003, pp. 22-24.

[20] HMM tool [Online]. Available:http://htk.eng.cam.ac.uk/

[21] Hollywood Dataset [Online]. Available: http://www.di.ens.fr/ laptev/download.html

[22] B. Hornler, ”Audio-Visual Event Recognition with Graphical Models”. Verlag Dr. Hut, 2010 (PhD Thesis).

[23] Q. Huang,and S. Cox. ”Inferring the structure of a tennis game using audio information.” , IEEE Transactions on Audio, Speech, and Language Processing vol. 19, no. 7, pp. 1925-1937, 2011

[24] jAudio tool [Online]. Available: http://jaudio.sourceforge.net/

[25] W. Jiang, C. Cotton, and A. C. Loui. ”Automatic consumer video summarization by audio and visual analysis.” IEEE International Conference on Multimedia and Expo (ICME), 2011, pp. 1-6.

[26] H. Jiang, T.Lin, and H. Zhang. ”Video segmentation with the support of audio segmentation and classification.” IEEE International Conference on Multimedia and Expo. 2000.

[27] Y. G. Jiang, S. Bhattacharya, S. Chang, and M. Shah. ”High-level event recognition in unconstrained videos.” International Journal of Multimedia Information Retrieval vol. 2, no. 2, 73-101,2013

[28] Q. Jin, P. S. F. Schulam, S. Rawat, S. Burger, D. Ding, and Florian Metze. ”Event-based Video Retrieval Using Audio.” In Interspeech 2012.

[29] Y. Kamishima, N.Inoue, and K. Shinoda. ”Event detection in consumer videos using GMM supervectors and SVMs.” EURASIP Journal on Image and Video Processing no. 1 , pp. 1-13,2013.

[30] B. Kedem Spectral analysis and discrimination by zero crossings IEEE Proceedings 1986, vol. 74 pp. 1477-1493.

[31] D. S. Kim, J. H. Jheong, J. W. Kim, S. Y. Lee, ”Feature extraction based on zero-crossings with peak amplitudes for robust speech recognition in noisy environments”, IEEE International Conference on Acoustics , Speech and Signal Processing, Volume 1, pp. 61-64, October 1996

[32] D. S. Kim, S. Y. Lee, and R. M. Kil ”Auditory processing of speech signals for robust speech recognition in real world noisy environments”, IEEE Transactions on Speech and Audio Processing vol. 7, no.1, pp. 55-69, January 1999.

[33] Kodak Benchmark dataset [Online]. Available: http://www.ee.columbia.edu/ln/dvmm/consumervideo/

[34] KTH dataset [Online]. Available: http://www.nada.kth.se/cvap/actions/

[35] K. Lee, and D. P.W. Ellis. ”Audio-based semantic concept classification for consumer video”, IEEE Transactions on Audio, Speech, and Language Processing vol.18, no.6, pp. 1406-1416, 2010.

[36] L. Ying, S. Narayanan, and C. C. J. Kuo. ”Content-based movie analysis and indexing based on audiovisual cues.” IEEE Transactions on Circuits and Systems for Video Technology, vol 14, no. 8, pp. 1073-1085, 2004

[37] Z. Liu, J. Huang, and Y. Wang. ”Classification of TV programs based on audio information using Hidden Markov Model.” IEEE Second Workshop on Multimedia Signal Processing, 1998.

[38] Z. Liu, J. Huang, and Y.Wang et al., ”Audio feature extraction and analysis for scene classification” Proc. IEEE 1st Multimedia Workshop, 1997.

[39] R. Lienhart, S. Pfeiffer, and W. Effelsberg. ”Scene determination based on video and audio features.” IEEE International Conference on In Multimedia Computing and Systems, 1999, vol 1, pp.685-690.

[40] A. C. Loui and et al., Kodak consumer video benchmark data set: concept definition and annotation, ACM Workshop on MIR, 2007.

[41] Mediaeval [Online]. Available: http://www.multimediaeval.org/about/

[42] D. Mitrovi, M. Zeppelzauer, and C. Breiteneder. ”Features for content based audio retrieval.” Advances in computers vol. 78, pp. 71-150, 2010.

[43] P. Muneesawang, T. Amin, and L. Guan. ”Audio visual cues for video indexing and retrieval.” Advances in Multimedia Information Processing-PCM , Springer Berlin Heidelberg, pp. 642-649, 2005.

[44] H.A Murthy and V. Gadde, ”The modified group delay function and its application to phoneme recognition”, In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, April 2003, vol 1, pp 68-71.

[45] T. Nagarajan, H. A. Murthy, ”Sub-Band based group delay segmentation of spontaneous speech into syllable-like units” EURASIP Journal on Applied Signal Processing, vol. 17 pp. 2614-2625, 2004

[46] Y. Nakajima, Y. Lu, M. Sugano, A. Yoneyama, H. Yamagihara, and A. Kurematsu. ”A fast audio classification from MPEG coded data.” IEEE International Conference on In Acoustics, Speech, and Signal Processing, 1999, vol. 6, pp. 3005-3008.

[47] [Online]. Available: http://opensmile.sourceforge.net/

[48] C. Penet, C. H. Demarty,G. Gravier, P. Gros, ”Audio Event Detection in Movies using Multiple Audio Words and Contextual Bayesian Networks.” CBMI-11th International Workshop on Content Based Multimedia Indexing, 2013.

[49] PRAAT tool http://www.fon.hum.uva.nl/praat/

[50] M. Roach, and J. S. D. Mason. ”Classification of video genre using audio.” In INTERSPEECH, 2001, pp. 2693-2696.

[51] Y. Rui, A. Gupta, and A. Acero. ”Automatically extracting highlights for TV baseball programs.” In Proceedings of the eighth ACM international conference on Multimedia, 2000, pp. 105-115.

[52] C. Saraceno and R. Leonardi. ”Audio as a support to scene change detection and characterization of video sequences.” IEEE International Conference on In Acoustics, Speech, and Signal Processing, 1997, vol.4, pp. 2597-2600.

[53] K. Shinoda, and N. Inoue. ”Reusing Speech Techniques for Video Semantic Indexing [Applications Corner].” IEEE Signal Processing Magazine, vol. 30, no.2, pp 118-122, 2013.

[54] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso. ”On the use of audio events for improving video scene segmentation.” In Analysis, Retrieval and Delivery of Multimedia Content, Springer New York, 2013.

[55] H. Sundaram, and S. Chang. ”Video scene segmentation using video and audio features.” IEEE International Conference on Multimedia and Expo, 2000, vol. 2, pp. 1145-1148.

[56] A. J. Stork, J. Silva, L. Spinello, O. Kai, G. D. Tipaldi, M. Luber. ”Audio-Based Human Activity Recognition with Robots” In International Conference on Social Robotics 2011, vol. 2.

[57] C. M. Taskiran, Z. Pizlo, A. Amir, D. Ponceleon, and E. J. Delp, ”Automated Video Program Summarization Using Speech Transcripts” IEEE Transactions On Multimedia, Vol. 8, No.4, August 2006.

[58] B. U. Treyin, Y. Dedeolu, and A. Enis etin. ”HMM based falling person detection using both audio and video.” In Computer Vision in Human- Computer Interaction, pp. 211-220, 2005.

[59] I. Trancoso, T. Pellegrini, J. Portelo, H. Meinedo, M. Bugalho, A. Abad, and J. Neto, ”Audio contributions to semantic video search.” IEEE International Conference on Multimedia and Expo, 2009, pp. 630-633.

[60] Trecvid Dataset [Online]. Available: http://trecvid.nist.gov/

[61] V. S. Tseng, J. Su, J. H. Huang and C. J. Chen. ”Integrated mining of visual features, speech features, and frequent patterns for semantic video annotation.” IEEE Transactions on Multimedia, vol. 10, no. 2, pp. 260-267, 2008.

[62] K. Umapathy, S. Krishnan and S.Jimaa, ”Multi group classification of audio signals using time frequency parameters”, IEEE Transactions on Multimedia vol.7,no. 2, pp. 308-315, April 2005

[63] J. Wang, C. Xu, E. Chng, and Q. Tian. ”Sports highlight detection from keyword sequences using HMM.” IEEE International Conference on In Multimedia and Expo, 2004, vol. 1, pp. 599-602.

[64] Z. Xiong, R. Radhakrishnan, A. Divakaran, and Thomas S. Huang. ”Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework.” IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003, vol. 5, pp. V-632.

[65] Z. Xiong, ”Audio-visual sports highlights extraction using coupled hidden markov models.” Pattern analysis and applications vol. 8, no.1-2, pp. 62-71, 2005.

[66] M Xu, N. C. Maddage, C. Xu. ”Creating audio keywords for event detection in soccer video.” IEEE International Conference on Multimedia and Expo, 2003, Vol. 2.

[67] M. Xu, C. Xu, L. Duan, J. S. Jin, and S. Luo. ”Audio keywords generation for sports video analysis.” ACM Transactions on Multimedia Computing, Communications, and Applications vol. 4, no. 2, pp. 11, 2008.

[68] Yaafe tool [Online]. Available: http://yaafe.sourceforge.net/

[69] Zhang.T. ”Automatic singer identification”, IEEE Conference on Multimedia and Expo, July 2003, Vol 1, pp. 33-36.

[70] B. Zhang,W. Dou, and L. Chen. ”Ball hit detection in table tennis games based on audio analysis.” 18th International Conference on Pattern Recognition, 2006, vol. 3, pp. 220-223.