A Survey on Audio Content-Based Classification

Conference Paper · December 2017

DOI: 10.1109/CSCI.2017.69

2 authors:

Amal Dandashi
College of North Atlantic Qatar
25 PUBLICATIONS **156** CITATIONS
SEE PROFILE

Jihad Mohamad Aljaam
Université de Technologie et de Sciences Appliquées Libano-Française
120 PUBLICATIONS **1,196** CITATIONS
SEE PROFILE

2017 International Conference on Computational Science and Computational Intelligence

A Survey on Audio Content-Based Classification

Amal Dandashi*, Jihad AlJaam
Department of Computer Science and Engineering
Qatar University
Doha, Qatar
{amal.dandashi;jaam}@qu.edu.qa

Abstract— **with the increased usage of mobile electronic devices, social media platforms, and electronic based applications in everyday life, the upload and usage of multimedia clips is exponentially increasing every year. There are many research initiatives tackling the challenge of multimodal video classification and it has been found that audio-based classification is less computationally expensive and just as effective, in many cases. However, research targeted towards acoustic-based detection is still in its initial stages. Audio content-based classification pertains to several domains: music and speech signal processing, which are relatively popular research interests, and event, genre and scene-based classification which are still areas that need a lot of development. There is also a problem with the difficulty of assessing performances of different systems with a unified audio dataset, due to the lack of development in this field. The objective of this study is to present information about audio processing techniques and studies conducted under several classification tracks; namely acoustic-based event, genre, and scene detection, as well as combination-based classification works.**

Keywords— audio processing; acoustic features; audio-based event detection, video classification.

I. INTRODUCTION (*HEADING 1*)

Events around the world have been widely escalating, and the advance of technology and social media has led to a vast proliferation in the presence of digital audio-visual content, particularly over the last decade. Citizen journalism has been rising, where citizens spontaneously document any event happening in their location, by utilizing smart phones to shoot videos and uploading them online. Much of this data is in raw form, unclassified, unannotated, and unused. Many research efforts have been directed towards the processing of audio-visual data. The predominant features in these systems include multimedia content exchange, open and shared delivery platforms that enable digital audio-visual (AV) content sharing between different users, semantic metadata retrieval and exploitation, and adaptive and personalized content discovery and delivery. However, the processing task of audio-visual content can be challenging and computationally expensive. In many cases, extracting audio from video and processing solely the audio modality on is less a computationally complex method to classify digital multimodal data.

The research community has contributed to increased interest in the field of audio processing; including segmentation, indexing, classification and retrieval of audio content. This is to due to the clear benefits of reproducibility of results, dissemination of open source audio processing code libraries, and documentations of results[1]. There have been various projects aimed at the public evaluation of audio-based systems, such as SiSEC evaluation for signal separation [2], MIREX for information retrieval in the music domain [3], and the CHiME challenge for speech separation and recognition [4]. Unfortunately for audio-based event classification, not involving speech or music, there is only the CLEAR evaluations [5], and the TRECVID challenge [6] which is focused on audio-visual multimodal event detection.

Automated audio event classification is an application of machine learning and pattern recognition where an audio signal is mapped to corresponding acoustic events in an auditory scene by use of a symbolic description. Audio-based event detection is utilized in various applications such as health care monitoring, surveillance, context-based indexing and retrieval in movies and sports videos, military applications, and audio segmentation. Accurate audio event detection is vital now with the rise of social media, video blogging and wearable electronics like Google Glass. Most audio event detection methods utilize Gaussian Mixture Models (GMMs), which operate with Mel-Cepstral Coefficients (MFCCs) [7,8]. Automatic Speech Recognition (ASR) techniques are also used for audio detection events, but they are not robust to background noise or simultaneous events[9,10] . Other challenges include the lack of abundance of training data that would be provided by audio event datasets which limits the ability to train accurate parametric models.

In this study, we present a survey of the state of the arts methods for a multitude of methodologies used for audio-based classification. The following section consists of an overview of audio processing information and most commonly utilized features. Section III depicts various studies conducted under five main classification schemes: audio-based event detection,

978-1-5386-2652-8/17 $31.00 © 2017 IEEE
DOI 10.1109/CSCI.2017.69

violence detection, genre detection, scene/environmental detection and combination based approaches. Section IV contains a description of several open source audio libraries. Section V includes several examples of standardized audio datasets developed specifically for acoustic detection and baseline testing. Finally, section VI concludes the study with a visit of the challenges faced in the audio processing research community.

II. Overview of Audio Processing

Audio-only approaches [11] are more commonly utilized than text only approaches for video classification. Audio approaches require fewer computational resources than that of visual methods. When features are stored, they also require less space. Another advantage is that segmented audio clips tend be very short (average 1-2 seconds), so the processing of the audio clips would be easier.

Audio features can lead to three layers of audio understanding: low-level acoustics, such as the average frequency for a frame, midlevel sound objects, such as the audio signature of the sound a ball makes while bouncing, and high-level scene classes, such as background music playing in certain types of video scenes.

Two main techniques are using either time domain features or frequency domain features. Using time domain means plotting amplitude of a signal with respect to time, while frequency domain means plotting amplitude with respect to frequency, which pertains to the spectrum of signal.

The volume standard deviation and volume dynamic range measure may be utilized for time domain features, i.e., sports has a nearly constant level of noise. Different classes of sounds may be categorized by setting certain thresholds. The zero crossing rate (ZCR) is the number of signal amplitude sign changes per frame. A high ZCR indicates high frequency, i.e., speech has a higher ZCR variability than music does. Silence ratio is the proportion of a frame with amplitude values measured with respect to some threshold, i.e., news has a higher silence ratio than commercials, and speech has a higher silence ratio than music.

Frequency domain suggests an energy (signal) distribution across frequency components. The frequency centroid approximates brightness, and is the midpoint of the spectral energy distribution, i.e., brightness is lower in speech than in music. Bandwidth is the measure of the frequency range of a signal, i.e., speech has lower bandwidth than music. The lowest frequency in a sample is the fundamental frequency, which approximates pitch, and may be used to distinguish between speaker genders, or to identify parts of speech such as introduction of a new topic. A frame that is not silent and does not have a pitch represents noise.

III. Audio content-based Classification Studies

There are several studies that have attempted video or multimedia classification using only audio signals. These methods can be classified based on different types of detection, such as music detection, genre detection, scene detection, event detection, emotion detection and others are more specific, like violence or hazardous circumstance detection. Many studies utilize audio in accordance with visual based classification: combination-based approaches. Others utilize deep learning techniques for audio-based detection. In this section, we present previous studies pertaining to audio-based event detection, genre detection, scene detection and some multimodal combination-based detection studies.

A. Audio Event Detection

The amount of user generated multimedia digital data on the internet has increased exponentially over the past decade. Among the most popular multimedia sites, YouTube, reported that 300 hours of digital recordings are uploaded every minute [12]. As of March 2015, there are 70+ million hours of watch time on YouTube. There are other popular internet sites which report similar statistics. The uploaded recordings are mostly unannotated, and descriptions are limited to high-level metadata, like author name or a brief title. Audio based event detection is vital for extracting descriptions of multimedia recording and content analysis of digital audio.

The authors in [13] propose a system framework for learning acoustic event detectors using only weakly labeled data. The study involves a demonstration of the problem being formulated as a Multiple Instance Learning problem. A two framework solution is then proposed for solving multiple-instance learning, one based on support vector machines (SVM) and the other on neural networks. The proposed approach leads to less time consuming and less expensive process of the manual annotation of data, in order to facilitate fully supervised learning. The system is able to recognize events and provide temporal locations of the events in the recordings. Results show that events like clanking, scraping and children’s voices are easily detectable using SVM and neural network approaches, whereas events such as drums, hammering and laughing are harder to detect using both of those methods.

Another study [14] deals with the detection of audio events derived from real life recordings. The authors develop a technique for detecting signature audio events based on identifying patterns of occurrences of automatically learned atomic units of sound, named Acoustic Unit Descriptors (AUDs). Experimental results demonstrate that the proposed methodology works well for individual event detection as well as their boundaries in complex recordings.

In this work [15], the authors present an exemplar-based method for audio based detection, based on non-negative matrix factorization (NMF) which is only considered in the context of audio event detection. Events are modeled as linear combinations of dictionary atoms, and mixtures as linear combination of overlapping events. The weights of the activated atoms serve as direct evidence for the underlying event classes. This eliminates the need for error-prone source separation after conventional audio event detection. The proposed work offers three main contributions: modelling training data through exemplars, artificially increasing the amount of training data via linear time warping of the spectra at various rates, and explicit modeling of background events like noise. Results yielded promising results on standardized datasets, however indicated problems with either overfitting and/or development test mismatches.

In this study [16], the authors lay out how the bag-of-words model commonly used for text or visual based classification has also been applied to audio-based classification; bag-of-audio words (BoAW). The proposed BoAW method extracts audio concepts in an unsupervised way. This gives it the advantage over other methods as it can be utilized easily for a new set of audio concepts in multimedia videos, without going through tedious manual annotation. Features are extracted from one-dimensional audio signals at fixed length intervals. These intervals may not capture the full acoustic variation that characterizes a specific sound. Experimental results depicted that certain representation decisions in the bag-of-visual-words algorithm such as L1-normalization are not optimal for audio representation. Results also varied in dependence on the acoustic variation of the video.

B. *Audio Violence Event Detection*

These audio detection systems centered on detection violence are conventionally developed in order to provide audio based surveillance to public areas, in order to prevent or detect crime. The factor that researchers focus on is to minimize the false alarm rate, in order to avoid unnecessarily alarming responsible personnel.

This study [17] proposes a technique for automatic space monitoring based solely on the perceived audio data. The main objective of this study is to detect abnormal hazardous events in a noisy background environment. The authors focus on events where dangerous situations take place in a metro station, such as screams, explosions and gunshots. The aim is to help warn authorized personnel to take precautions or take actions to prevent crime and property damage. In order to do this, the false alarm rate must be to a minimum. The approach utilized is based on a two stage recognition schema, which both utilize HMMs and GMMs to extract the approximate density function of the corresponding acoustic class. The feature set used was MFCC augment with a second group of parameters based on the MPEG-7 audio standard. Performance evaluation reports high detection rates in terms of false alarm and miss probability rates.

Another study [18] is centered on audio classification specifically for events concerning citizen security in urban environments. The various events studied are: explosion, broken glass, shot, shout and others. The objective of this study is to build and test a system that performs audio-event detection for these specific danger situations, using MFCC features and HMM-based representation of acoustic data. The system is trained of a dataset of recordings developed by the authors. Performance results achieved promising results but could stand to use much optimization by tweaking the feature parameters.

In another study [19], authors present a method of violent shot detection in movies. They utilize audio and video modalities to classify, separately at first, and combine them at the end. For audio-based detection, a weakly-supervised method is used to improve classification accuracy, to detect whether the movie is violent or non-violent. Then they tackle detecting the violent event more specifically using visual-based classification to detect motion, flame, explosion and blood related events. Probabilistic Latent Semantic Analysis (PLSA) is utilized for this technique, and they test results on five movies, comparing the results of PLSA with the SVM classifier. The authors found enhanced results with their technique.

C. *Audio Genre Detection*

Lui et al. [20] used sample audio signals at a specific frequency, and after segmenting and subdividing into overlapping frames, utilized the following audio features: nonsilence ratio, volume standard deviation, volume dynamic range, pitch standard deviation, and others. Results depicted that the features with the highest discriminatory power are frequency centroid, frequency bandwidth, and energy ratio. Classification was then performed using one-class-one-network structure. The audio samples were then classed into commercial, basketball, football, news report, and weather forecast categories.

Roach and Mason [21] have utilized audio from video for the purpose of genre classification. They used Mel-frequency cepstral coefficients which are coefficients derived from a cepstral representation of an audio clip. This approach was utilized due to its success with speech recognition. The authors find that best results are achieved with 10-12 coefficients. Classification is performed with the Gaussian mixture model due to its effectivity for speaker recognition. The genres studied are fast-moving sports, cartoons, news, commercials and music.

Dinh et al. [22] use a Daubechies 4 wavelet to seven sub-bands of TV show audio clips. Wavelet transforms are useful for reducing dimensionality and have good energy compaction. The audio features used are sub-band energy, subband variance, zero crossing rate, as well as two customized features; centroid and bandwidth. Classifiers used are the C4.5 decision tree, K nearest neighbor, and support vector machine. Clips of different lengths not higher than 2 seconds were tested, and depicted no significant difference in performance. The genres tested were vocal music shows, news, commercials, cartoons and motor racing sports.

Moncrief et al. [23] utilize audio-based cinematic principles to distinguish between horror and nonhorror films. Variations in energy intensity were used to detect sound energy levels, which in this study are associated with feelings of surprise, alarm, apprehension, surprise followed by alarm, and apprehension progression to climax. These four types of sound were found to be effective to distinguish horror movies and even to distinguish scenes within a horror movie.

D. *Audio Scene Detection*

Although the majority of studies surrounding acoustic classification have conventionally focused on music and speech signal processing, the challenge of acoustic environmental or scene detection has received more attention over the past few years. Recent work has focused more on non-stationary aspects of scenic sounds, and various new features centered on that have been proposed. In addition to that, sequential learning methods have been used to account for long term variation of environmental sounds.

This study [24] presents a challenge on the detection and classification of acoustic scenes and events. The authors ran a scene classification challenge, and two event detection and classification challenges, namely the office live (OL) and office synthetic (OS). The objective was to highlight areas that need improvement, to the research community concerned with audio-based scene detection. Results depicted that, in the case of scene classification, simple systems can do relatively well, however complex systems can bring performance to the levels achieved by human listeners. The strongest performers chose a diverse set of features, used temporal information, and often used SVMs for classification.

In this study [25], a survey is conducted targeting acoustic scene detection studies. This work is centered on three main themes: basic environmental sound processing methods, stationary techniques, and non-stationary techniques. Stationary techniques are dominated by spectral features, which are easy to compute but have limitations in the modeling on non-stationary sounds. Non-stationary techniques obtain features pertaining to the wavelet transform, the sparse representation and the spectrogram. The latter two get the best results for non-stationary environmental sound detection. MFCC features are also utilized, often in combination with several other features to boost classification accuracy. Non-stationary methods give the best results, but are the most computationally expensive.

The authors also point out that each paper in this field presented its performance evaluations with their own datasets, due to a lack of standard datasets available for testing. This makes it difficult to conduct a fair quantitative comparison of different approaches.

E. Combination Based Approaches

Many studies incorporate the use of several combinations [26] of text, audio and visual features in order to complement each technique and overcome weaknesses of each. The main challenge of utilizing features from different modalities is knowing how and when to combine these features[27].

Qi et al. [28] use audio, visual and textual features to classify news streams into genres of news stories. Audio and visual features are utilized to segment and group video shots into scenes. Text processing is used after detection of text through closed captions or scene text detection. Support vector machine classifier is used to classify the news stories.

Jasinschi and Louie [29] classify TV shows using audio, visual and textual features. The audio features are used to classify six categories; noise, speech, music, speech and noise, speech and speech, speech and music. Visual features are utilized to detect commercials. Textual features segment noncommercial parts of the TV program via annotations in closed captions. Finally all audio categories are combined to classify the TV program as financial news or talk show.

Roach et al. [30] extend on their previous work, which consisted of classifying videos using audio features, to include using visual features. Adapted Gaussian Models for Image Classification (AGMM) is the classifier used for linear combination of the conditional probabilities of visual and audio features. The video classes studied are news, commercial, sports, cartoons and music videos.

Rasheed and Shah [31] utilize cinematic principles with accordance to audio and visual features to classify movies by analyzing the movie previews. Intersection of hue, saturation and value (HSV) color histograms are used to segment previews into shots. Motion per preview is then calculated by using the ratio of moving pixels to total pixels per frame (visual disturbance), for each frame per preview. After visual disturbance is plotted against average shot length, a linear classifier is used to distinguish action and nonaction movies. Then audio energy variation analysis is used to categorize action movies into those with fire or explosions, or without. Light intensity thresholds are used to classify movies as comedy, drama or horror. Horror movies have low levels of light intensity while comedies have the highest light levels, and dramas are in the middle.

IV. Audio Open Source Code Libraries

There are several open source software projects dedicated to progressing the audio processing/classification research community. The open source audio feature extraction toolbox, Yaafe [32] consists of audio-processing tools that use an audio-based combination of features to achieve a statistical learning model, in order for future events to be classified. The Yaafe toolbox includes several intermediate representations such as spectrum, envelope and autocorrelation. It also includes options for temporal integration. There are several studies which have utilized the Yaafe toolset for audio-based component to classify videos and music[33,34].

There are other audio processing libraries built on the base of Yaafe features such as Essentia [35], which consists of audio processing features, and works with the GAIA plugin that complements the ESSENTIA studio with classifiers. The ESSENTIA project is an open source C++ library for audio analysis and audio-based music data retrieval. Its algorithm collection includes audio input/output functionalities, statistical characterization of data, digital signal processing blocks. The features provided in ESSENTIA include spectral, temporal, tonal and high-level music descriptors.

V. Audio Datasets

An audio classification system must be tested on standardized audio datasets in order to evaluate performance results, in comparison with similar studies. Standardized audio databases developed for testing are limited in number. We present here an overview on several developed audio datasets utilized for baseline testing.

The TUT Acoustic Scenes 2016 database [36] has been developed, for environmental audio research. It consists of binaural recordings from 15 different acoustic environments, indoor and outdoor. TUT Sound Events 2016, a subset of this dataset contains annotations for specific sound events. It consists of residential and home environments and is manually annotated. The authors provide a description of the database content, the recording and annotation procedure, along with a protocol for cross validation and setup and performance results of acoustic scene classification and event detection, with the use of MFCC and GMMs. The TUT Acoustic Scenes 2016 dataset includes 15 acoustic scenes, namely: bus, beach, restaurant, city center, grocery store, home, library, forest path, car, metro station, train, tram, park, residential area and home. All audio segments are 30 seconds long.

Acoustic scene classification pertains to the recognition of the audio environment, with applications in technology requiring environmental awareness. The environment may be specified in terms of physical or social context, for example, park, house, office, meeting, etc. Other databases for acoustic scene development include the DCASE 2013 [37], and the LITIS Rouen Audio Scene dataset [38].

The 2010 community-based Signal Separation Evaluation Campaign (SiSEC2010) includes an audio dataset developed specifically for the task of baseline testing for speech and music audio-based classification [39]. It contains seven speech and music datasets, including datasets recorded in noisy or dynamic environments, along with the SiSEC 2008 datasets. The authors provide a protocol for testing of five main tasks, and an evaluation guide using different objective performance criteria.

Another database that targets speech processing studies and evaluations is the Open-Source Multi-Language Audio Database for Spoken Language Processing Applications [40] which contains speech passages from YouTube, specifically, 300 passages in three languages; English, Mandarine and Russian. The Multichannel audio database [41] consists of an acoustic dataset with audio segments in various acoustic environments designed in order to measure impulse response.

VI. Conclusions

With the age of the internet, increased usage of smart electronic devices and the exponential growth of social platforms for individual expression, digital multimedia has been increasingly uploaded in raw and unannotated form online. The need for technology systems to accurately classify and detect audio and multimedia has never been so vital. In addition to the need for classification, retrieval and reuse of user-uploaded audio clips, there are many other real-life applications that could stand to benefit from digital audio automated classification, some of those being music/movie platforms which could take advantage of audio-based genre classification, bank, surveillance and security applications that could benefit from acoustic scene or speech classification, hospital based monitoring, military applications and many others.

This study sheds light on information about the techniques and methodologies of audio classification, and presents a multitude of various studies conducted, with a focus on acoustic event, genre, scene, and combination-based classification. We also present various audio databases developed for baseline testing in specific fields. The aim of this paper is to highlight areas needed for improvement and future work within the field of audio-based classification. There are many studies centered on acoustic music and speech signal detection. Features such as zero-crossing rate, short-time energy, and spectrum flux are mainly used in the speech and music classification fields, whereas features like band periodicity and noise frame ratio are mainly used for scene/environment or music based classification. Linear spectrum pairs and MFCCs are utilized in all forms of classifications. Audio-based event and scene classification research is still in the initial research and development stages, and could stand to be enhanced with optimized algorithms, feature combinations, and improved classifiers In addition to that, the lack of standardized audio datasets makes it a difficult task to evaluate and compare results between research communities.

Acknowledgment

This publication was made possible by GSRA grant # 1-1-1202-13026 from the Qatar National Research Fund (a member of Qatar Foundation). The findings achieved herein are solely the responsibility of the author(s).

References

[1] P. Vandewalle, J. Kovacevic, and M. Vetterli. "Reproducible research in signal processing." *IEEE Signal Processing Magazine* 26.3 (2009).

[2] G. Nolte, et al. "The 2011 Signal Separation Evaluation Campaign (SiSEC2011):-Biomedical Data Analysis-." *LVA/ICA*. 2012.

[3] “Music Information Retrieval Evaluation eXchange (MIREX),” http://music-ir.org/mirexwiki/.

[4] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, “The PASCAL CHiME speech separation and recognition challenge,” *Computer Speech & Language*, vol. 27, no. 3, pp.621–633, May 2012.

[5] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan, “The CLEAR 2006 evaluation,”Multimodal Technologies for Perception of Humans, pp. 1–44, 2007.

[6] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A. F. Smeaton, and G. Quenot, “TRECVID 2012 – an overview of the goals, tasks, data, evaluation mechanisms and metrics,” *in Proc. of TRECVID 2012*. NIST, USA, 2012.

[7] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, “Real-world acoustic event detection,” *Pattern Recognition Letters,* vol. 31, no. 12, pp. 1543 – 1551, 2010.

[8] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, “Context dependent sound event detection,” *EURASIP Journal on Audio, Speech, and Music Processing*, 2013.

[9] J. F. Gemmeke, A. Hurmalainen, T. Virtanen, and Y. Sun, “Toward a practical implementation of exemplar-based noise robust ASR,” in *Proc. EUSIPCO*, 2011, pp. 1490–1494.

[10] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, “Exemplarbased sparse representations for noise robust automatic speech recognition,” *IEEE Transactions on Audio, Speech and Language processing,* vol. 19, no. 7, pp. 2067–2080, 2011.

[11] M. H. Lee, S. Nepal, and U. Srinivasan, Edge-based semantic classification of sports video sequences, in *Proceedings of the International Conference on Multimedia and Expo,* vol. 2, pp. 157–160, 2003.

[12] Youtube statistics.http://www.youtube.com/yt/press/statistics.html.

[13] A. Kumar, and R. Bhiksha. "Audio event detection using weakly labeled data." *In Proceedings of the 2016 ACM on Multimedia Conference*, pp. 1038-1047. ACM, 2016.

[14] A. Kumar, P. Dighe, R. Singh, S. Chaudhuri, and B. Raj. "Audio event detection from acoustic unit occurrence patterns." *In Acoustics, Speech and Signal Processing (ICASSP*), 2012 IEEE International Conference on, pp. 489-492. IEEE, 2012.

[15] J. Gemmeke, L. Vuegen, P. Karsmakers, and B. Vanrumste. "An exemplar-based NMF approach to audio event detection." *In Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013 IEEE Workshop on, pp. 1-4. IEEE, 2013.

[16] S. Pancoast, and M. Akbacak. "Bag-of-audio-words approach for multimedia event classification." *In Thirteenth Annual Conference of the International Speech Communication Association*. 2012.

[17] S. Ntalampiras, I. Potamitis, and N. Fakotakis. "On acoustic surveillance of hazardous situations." *In Acoustics, Speech and Signal Processing,* 2009. ICASSP 2009. IEEE International Conference on, pp. 165-168. IEEE, 2009.

[18] M. Pleva, E. Vozáriková, S. Ondáš, J. Juhár, and A. Čižmár. "Automatic detection of audio events indicating threats." *In IEEE International Conference on Multimedia Communications, Services and Security*, Krakow, vol. 6, no. 7.5. 2010.

[19] J. Lin, and W. Wang. "Weakly-supervised violence detection in movies with audio and video based co-training." *Advances in Multimedia Information Processing-PCM* 2009(2009): 930-935.

[20] Z. Liu, J. Huang, and Y. Wang, Classification of TV programs based on audio information using hidden Markov model, in *Proceedings of the IEEE Multimedia Signal Processing Workshop ,* pp. 27–32, 1998.

[21] M. Roach and J. Mason, Classification of video genre using audio, In *Interspeech*, vol. 4, pp. 2693–2696, 2001.

[22] J.-Y. Pan and C. Faloutsos, Videocube: A novel tool for video mining and classification, In *International Conference on Asian Digital Libraries*, pp. 194-205, Singapore, 2002.

[23] S. Moncrieff, S. Venkatesh, and C. Dorai, Horror film genre typing and scene labeling via audio analysis, In *Proceedings of the International Conference on Multimedia and Expo,* vol. 1, pp. 193–196, 2003.

[24] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M.D. Plumbley. "Detection and classification of acoustic scenes and events: An IEEE AASP challenge." *In Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013 IEEE Workshop on, pp. 1-4. IEEE, 2013.

[25] S. Chachada, and C-C. Jay Kuo. "Environmental sound recognition: A survey." *In Signal and Information Processing Association Annual Summit and Conference* (APSIPA), 2013 Asia-Pacific, pp. 1-9. IEEE, 2013.

[26] D. Brezeale, and D. J. Cook, Automatic video classification: A survey of the literature, In *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on,* 38(3), 416-430, 2008.

[27] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, Multimodal fusion for multimedia analysis: a survey, In *Multimedia systems*, 16(6), 345-379, 2010

[28] W. Qi, L. Gu, H. Jiang, X.-R. Chen, and H.-J. Zhang, Integrating visual, audio and text analysis for news video, In *Proceedings of the 7th IEEE International Conference on Image Processing (ICIP)*, pp. 520–523, September 2000.

[29] R. S. Jasinschi and J. Louie, Automatic TV program genre classification based on audio patterns, In *Proceedings of the IEEE 27th Euromicro Conference*, pp. 370–375, 2001.

[30] M. Roach, J. Mason, and L.-Q. Xu, Video genre verification using both acoustic and visual modes, In *International Workshop of Multimedia Signal Processing*, pp. 157–160, 2002.

[31] Z. Rasheed and M. Shah, Movie genre classification by exploiting audiovisual features of previews, In the *IEEE International Conference of Pattern Recognition*, vol. 2, pp. 1086–1089, 2002.

[32] B. Mathieu, et al., YAAFE, an Easy to Use and Efficient Audio Feature Extraction Software, In *Proceedings of the 11th International Conf erence on Music Information Retrieval (ISMIR 2010).* 2010.

[33] C. Frisson, et al., Videocycle: user-friendly navigation by similarity in video databases, In *Advances in Multimedia Modeling*. Springer Berlin Heidelberg, pp. 550-553. 2013.

[34] C. Copeland, and S. Mehrotra, Musical Instrument Modeling and Classification.

[35] D. Bogdanov, et al., ESSENTIA: an open-source library for sound and music analysis, In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013.

[36] A. Mesaros, T. Heittola, and T. Virtanen. "TUT database for acoustic scene classification and sound event detection." *In Signal Processing Conference (EUSIPCO)*, 2016 24th European, pp. 1128-1132. IEEE, 2016.

[37] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M.D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, Oct 2015.

[38] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of timefrequency representations for audio scene detection," *Tech. Rep., HAL*, 2014.

[39] S. Araki, A. Ozerov, V. Gowreesunker, H. Sawada, F. Theis, G. Nolte, D. Lutter, and N. Duong. "The 2010 signal separation evaluation campaign (SiSEC2010): Audio source separation." *In International Conference on Latent Variable Analysis and Signal Separation*, pp. 114-122. Springer, Berlin, Heidelberg, 2010.

[40] S. Zahorian. "Open-source multi-language audio database for spoken language processing applications." *STATE UNIV OF NEW YORK AT BINGHAMTON DEPT OF ELECTRICAL AND COMPUTER ENGINEERING*, 2012.

[41] E. Hadad, F. Heese, P. Vary, and S. Gannot. "Multichannel audio database in various acoustic environments." *In Acoustic Signal Enhancement (IWAENC*), 2014 14th International Workshop on, pp. 313-317. IEEE, 2014.