National Conference on Recent Research in Engineering and Technology (NCRRET-2015)
International Journal of Advance Engineering and Research Development (IJAERD)
e-ISSN: 2348 - 4470 , print-ISSN:2348-6406

A REVIEW ON AUDIO FEATURES BASED EXTRACTION OF SONGS FROM MOVIES

Mittal C. Darji[1],Dr. Narendra M. Patel[2], Zankhana H. Shah[3]
Birla Vishvakarma Mahavidyalaya, Vallabh Vidyanagar, Anand, Gujarat, India
mittaldarji54@gmail.com, nmpatel@bvmengineering.ac.in, zankhana.shah@bvmengineering.ac.in

ABSTRACT
A movie is a collection of different portions like comedy scenes, action, dialogs and songs. Extraction of a selected portion such as songs is a complex task. Basic three approaches namely visual based, text based and audio based are used by authors over the time for video content classification. This paper indicates that audio based approach is most suitable for extraction of songs. Audio based approach utilizes various audio features which can differentiate songs and non-songs as mentioned here. Different combinations of these features have been used by researchers from which a generic model is derived that can become the base for songs extraction application.

Index Terms: Video Classification, Audio Features, Segmentation, Time Domain, Frequency Domain, Zero-Crossing Rate, Silence,Tempo, Short Time Energy, Intensity, Bandwidth

1. INTRODUCTION

Production of the multi-media data is increasing with the time. There is a huge amount of content available for users including sports, news, documents, pictures, TV shows, presentations etc. searching of desirable data from this big data is a very time consuming and difficult task. Video segmentation and classification came into idea which works on different informative dimensions of a video namely audio, visual and text. Various methods have been found out in audio based approach, visual based approach and text based approach according to the applications to classify videos.

A journey from black and white silent film to a high-definition colored film has led to a tremendous amount of movies out there for viewers.Movie is a collection of various scenes like comedy, action, dialogues, drama and songs. Viewer would like to watch only a selected portion such as songs from this. The revenue generated by the music and songs of a movie is around 4-5% of the total. [2] The importance of film music and songs can be realized as well from the fact that 48% of India's music sales are from film music. [3]

Extraction of songs from movies can be beneficial for differ users like producers, dancers, singers or CD/cassette seller. But manually doing so is a very tedious, complex and time consuming task. Taking into account this need and the huge data out there, attempts have been made to develop the system which can perform this task automatically and accurately. Authors have used various approaches to separate songs from videos.

This review paper comparesthree approaches of classification which show that audio based approach is well suited for extraction of songs. Also, a review is made on various audio features which can help to differentiate songs from non-song. At the end a generic model is derived from the survey which can be taken as a base for development of an application to extract songs from movies.

Rest of the paper is organized as follows. In section 2 three classification approaches are reviewed. Section 3 describes audio features with their characteristics regarding songs and non-songs. Section 4 derives the generic model for song extraction from the survey of attempts made by researchers. Finally conclusion is provided in section 5 followed by the future work in section 6.

2. CLASSIFICATION APPROACHES

Video contains three types of data: (1) Visual (2) Text (3) Audio. Video classification can be performed using any of these three dimensions. Many applications are proposed in literature where classification techniques are applied. One approach can be suitable for some application while it might be less suitable for another application. This paper tried to find the most suitable classification approach for the application of extracting songs from movies. For that a survey of all three approaches is done here.

2.1 Visual based approach: Visual is the main and large information in a video and is used by many authors for successful partitioning of video. Looking inside the structure of a video, it contains a sequence of images known as frames. A sequence of interrelated continuous frames taken from a single camera which represents a single action in time is known as shot. Many of such shots make a scene. Collection of such scenes is a video.
This structure of video is utilized in visual approach for classification. Features of frames or shots are calculated and then their differences and similarities are used by researchers. Methods like shot based approach, hidden markov model technique, histogram based detection, graph theoretic dominant set approach are used.

2.2 Text based approach: Text-only approaches are the least common in the video classification literature. [7] Video contains textual data of two categories: (1) text displayed on screen in video (2) dialogs extracted form speech. Text from the screen such as name of a building is extracted by identifying the object followed by Optical Character Recognition (OCR) method to extract text from object. Dialogs can be extracted from speech by speech recognition methods or from subtitles by OCR. However subtitles do not give clear information about the sounds.

2.3 Audio based approach: Audio based approaches are found more often than text based approaches in video classification. Audio is an important part of a movie and contains a large amount of information which can be utilized in various ways to segment video. Audio part possesses various properties like intensity, pitch, Zero Crossing Rates (ZCR), Mel-Frequency Cepstral Coefficients (MFCC), spectrum flux, Root Mean Square (RMS), tempo, beat and chroma. Various combinations of these properties are applied in literature.

Table 1. Comparison of classification approaches

Approach	Advantage/Disadvantage	Applications
Visual Based	Large size of data Pre-processing or normalization is required due to lighting effect, in plane or out of plane rotation or object motion Overall greater computational resources and time required	Separating news reader and sight scenes from a news video Separating or tracking object Highlight extraction from sports Video summarization
Text Based	Speech recognition has comparatively high error rates OCR is computationally expensive	Reading score board Reading headlines from news video Providing subtitles for deaf people Language translation
Audio Based	Requires fewer computational resources than visual methods If features need to be stored then audio features require less space Audio clips are shorter in length and smaller in size than video clips	Separating songs, environmental noise, dialogs from movies Classifying videos into genre like news, commercials and sports. Classifying movies into genre like horror, non-horror or action.

A review of all three approaches with respect to their suitable applications and performance is given in Table 1. **From the table it is clear that audio based approach is best suitable for extraction of songs from movies.**

3. AUDIO FEATURES

Audio signal has many features which can distinguish song audio clip from the other part of movie. Samples taken from an audio clip of song are plotted in Figure 1. Same way audio clip of a non-song is shown in figure 2. Both the clips are of length 20 seconds. It is clearly observed from both the figures that song has overall higher intensity and lesser gaps or silence in between. Non-song possesses greater amount of silence. This is due to the small pauses in between the words while speaking. These figures also help in observation of few other audio features from table 2.

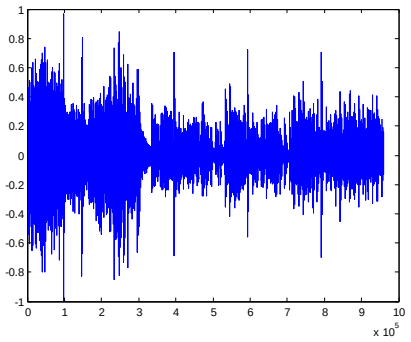

Fig 1. Song sequence of 20 sec

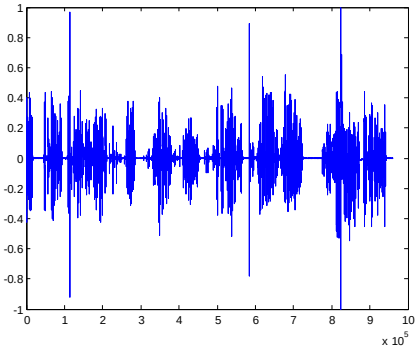

Fig 2. Non-song sequence of 20 sec

Figure 1 and 2 displays song and non-song audio signals respectively.

Audio features can be classified mainly in two types: (1) time domain features and (2) frequency domain features. Both the types of features are used by researchers.Time Domain Features like Root Mean Square (RMS) of energy or amplitude, Short Time Energy (STE), intensity, Zero Crossing Rates (ZCR) and silence are used. Time domain signal can be converted into frequency domain using fourier transform. It is also known as spectrum of a signal. Frequency domain features like Mel-frequency Cepstral Coefficients (MFCC), pitch, tempo and bandwidth are used.

Characteristics of these features differ for song and non-song which is mentioned in table 2 with the description of features.

Table 2. Characteristics of Audio features

Audio feature	Description	Song	Non-song
Root Mean Square	It is the RMS value of energy or amplitude of signal.	High	Low
Intensity	It is the sound power per unit area.	High	Low
Zero Crossing Rates	It is the rate of sign-changes along a signal. It is defined as the number of times zero crossed within a frame.	High	Low
Silence	It is the proportion of a frame with amplitude values below some threshold.	Low	High
Tempo	It is a musical terminology which describes the speed or pace of a given audio. Beats per minute (BPM) is a typically used unit for measurement of tempo.	Constant	Varying
Spectrum flux	It is the spectrum value difference between two adjacent frames.	Low	High
Short Time Energy variations	It represents the total power spectrum of a frame.	Low	High
Bandwidth	It is a measure of the frequency range of a signal.	High	Low

Above table shows how audio features can distinguish songs from non-song in movies.

4. SONG EXTRACTION FROM MOVIES

Many attempts have been made in literature where video is classified into categories like speech, music, environmental sounds and silence using only audio features.To separate song portion, comparison of audio feature values is required among parts of movie and hence segmentation of movie audio clip into smaller clips is the first task performed by most of the authors. Researchers have selected various combinations of features and segmentation ways in literature.

T. Ratanpara, M. Bhatt [2] has used Intensity as the major discriminator feature after the segmentation process. But using only intensity did not give correct result and hence they have used tempo and silence as well to remove the faulty detection of songs.

S. M. Doudpota, S. Guha [3] performed segmentation of movie and used three features at first stage namely ZCR, spectrum flux and STE. They also got false detection after first stage. Elimination of sequences having length less than 120 seconds is performed. Authors have generated the Probabilistic timed automata using the song grammar to accurately extract songs from movies.

C. Panagiotakis and G. Tziritas [4] made attempt to separate speech and music form videos. They performed real time segmentation using the RMS of amplitude. Silence and ZCR are used here for improvements in result.

H. Jiang, T. Lin, H. Zhang [5] classified video into parts like speech, music, environmental sounds and silence. Features like ZCR, STE and noise frame ratio were used. They used color histogram for corrections.

H. Harb, L. Chen, J. Auloge [6] separated speech, music and silence from the videos. For this, energy, ZCR and silence are used. Further they have performed gender detection using frequency and ZCR.

From the survey of all such researches it is clearly observed that only one audio feature is not sufficient to extract songs from movies. Using only one feature leads to promising songs which may have non-song portions as well. Usage of more than one audio feature is required for removal of such non-songs portion and improvement in accuracy in results. A general methodology is derived for extraction of songs form movies which is presented in figure 3.

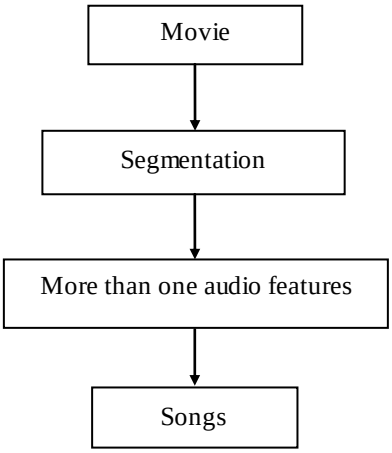

Fig 3. Generic model for song extraction

Figure 3 displays the methodology used by many authors over the time.

5. CONCLUSION

The work in this paper tried to review video classification approaches namely visual based approach, text based approach and audio based approach. From which it has been found out that audio based approach is most suitable for application that extract songs from movies.

Various audio based features have been surveyed here which can help to differentiate songs from non-songs. Different combinations of such features are used by authors over the time. Review of such literature provided important remarks: (1) segmentation is the very first step which is required for comparison of feature values (2) use of only one audio feature is not sufficient to accurately extract songs from movies. Using these remarks a generic model has been derived which can serve as a base for development of an application to extract songs from movies.

6. FUTURE WORK

Detailed designing or methodology for the blocks of generic model can be provided in future. An efficient segmentation method along with a good combination of audio features will make an accurate song extraction application.

7. REFERENCES

[1] S. Bhimani, A. Revar, A. Bhimani, "RMS Based Video Song Sequence Extraction Using Continuity Rule from Bollywood Movies", CIIT, vol. 6, No 9, 2014.

[2] T. Ratanpara, M. Bhatt ,"A novel approach to retrieve video song using continuity of audio segments from Bollywood movies", ITSIP & CIIT, October 2013.

[3] S. M. Doudpota, S. Guha,"Mining Movies to Extract Song Sequences", ACM 978-1-4503-0841-0 MDMKDD'11, August 21, 2011.

[4] C. Panagiotakis and G. Tziritas,"A speech/music discriminator based on rms and zero-crossings.", IEEE Transactions on Multimedia, 7:155–166, 2005.

[5] H. Jiang, T. Lin, H. Zhang, "Video segmentation with the support of audio segmentation and classification" Microsoft Research, China.

[6] H. Harb, L. Chen, J. Auloge ," Speech/Music/Silence and Gender Detection Algorithm" 7th International conference on Distributed Multimedia Systems DMS01, 2001.

[7] D. Brezeale, D. Cook, "Automatic Video Classification: A Survey of the Literature", IEEE Transactions in Volume:38, Issue: 3, 2008.

[8] C. V. Jawahar, B. Chennupati, B. Paluri, N. Jammalamadaka, "Video Retrieval Based on Textual Queries", Proceedings of the Thirteenth International Conference on Advanced Computing and Communications, Coimbatore, December 2005.

[9] V. T. Chasanis , A. C. Likas, and N. P. Galatsanos, "Scence Detecetion in videos using shot clustering and sequence Alignment", IEEE transactions on multimedia,vol 1. January 2009.

[10] P. Geetha, V. Narayanan, "A Survey of Content-Based Video Retrieval", Journal of Computer Science 4 (6): 474-486, 2008.

[11] Carey, M.J., Chepstow, L. Thomas, "A comparison of features for speech, music discrimination" Acoustics, Speech, and Signal Processing, Proceedings, IEEE International Conference in vol. 1, 1999.

[12] E Wold T. Blum, D. Keislar, and J. Wheaton,"Content-based classification, search, and retrieval of audio,", IEEE MultiMedia, vol. 3, no. 3, pp. 27–36, 1996.

[13] Lie Lu, Stan Z. Li and Hong-Jiang Zhang, "Content-based audio segmentation using support vector machines", Multimedia and Expo, ICME, IEEE International Conference on 2001.

[14] K. El-Maleh, M. Klein, G. Petrucci, P. Kabal, "Speech/Music discrimination for multimedia applications", Acoustics, Speech, and Signal Processing, ICASSP, Proceedings, IEEE International Conference in vol. 6 2000.

[15] J. Ajmera, I. A. McCowan, H. Bourlard, "Robust hmm-based speech/music segmentation", Acoustics, Speech, and Signal Processing (ICASSP), IEEE International Conference in vol. 1 2000.

[16] R. Natarajan, S. Chandrakala, "Audio-Based Event Detection in Videos - a Comprehensive Survey", International Journal of Engineering and Technology (IJET), Vol 6 No 4 Aug-Sep 2014.

[17] Information on Tempo available: http://en.wikipedia.org/wiki/Tempo

[18] Information on Zero-Crossing Rates available: http://en.wikipedia.org/wiki/Zero-crossing_rate