ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Music and Mood Detection using BoF Approach

Harshali Nemade [1], Deipali Gore [2]

PG Student, Dept. of Comp. Engineering, PESs MCOE, Savitribai Phule Pune University, Pune, Maharashtra, India

Asst. Prof., Dept. of Comp. Engineering, PESs MCOE, Savitribai Phule Pune University, Pune, Maharashtra, India

ABSTRACT: Music Information Retrieval (MIR) is the task of retrieving information from the music and it is the fastest growing area of music industry. Music Information Retrieval basically deals with the problems of querying and retrieving certain types of information from audio files with the help of large data set. Digital music is widely available in different digital formats due to explosive growth of information and multimedia technologies. Thus, the management and retrieval of such music is necessary for accessing music according to their meanings in respective songs. A lot of research and study has been going on in the field of music genre classification and music mood detection in the recent years. The basic approach of the work presented in this paper is for automatic identification of genre and mood detection of underlying audio file by mining different audio features. BoF representation is the way to represent the complex and high dimensional audio data. In the bag-of-frames approach, the signal is represented as the long-term distribution of the 'local'(frame-based) acoustic features. An early and late temporal pooling concept of sparse coding has been used to perform the classification tasks in more efficient way to improve the accuracy of the system.

KEYWORDS: Music Information Retrieval; BoF representation; Encoding; Pooling; Mood models; audio features.

I. INTRODUCTION

A. Music and mood

Listening music is one of the oldest and easiest way of entertainment for human beings. It is one of the basic human needs for recreation and entertainment. Studies have shown that listening right music at right time helps in healing, refreshing as well as inspiring human mind in any difficult situations. The music in the olden days was limited to live concerts, Performances or radio broadcasts is now available at everyone's finger tips within few clicks. Music is thus become very easily accessible and available. Music Information Retrieval (MIR) systems aims at extracting information from music. MIR deals with searching and retrieving audio files in efficient and effective manner. The overall collection of songs is nearly few millions, but we do not listen to same type of song again and again instead our choices differ time to time. We can say that we have our own preferences depending on various factors like music type, composer, artist, singer, instrument, location, albums etc. However, choosing a song or musical piece suiting our mood from a large collection is difficult and time consuming, since each of the mentioned parameter cannot sufficiently convey the emotional aspect associated with the song. Categorizing music as per its mood is a harder task because detecting a category to which mood will belong to has following challenges:-
1] Properties of music affect the emotions of music. Human mood, surrounding environment, cultural background ,education , individual personality, choices of songs etc affects the categorization of music and providing some label to the song as most of the research involves listeners who judges the emotional meaning of piece of song.
2] Adjectives describing emotions can be ambiguous.

B. Music Information Retrieval

Music Information Retrieval (MIR) is the science of retrieving information from the musical piece. MIR is useful for searching and retrieving appropriate information of respective music from a large musical dataset. This retrieved information can be used in many applications such as recommender systems, instrument recognition, artist recognition, musical annotation, separation applications, automatic categorization systems, speech recognition, environmental sound recognition and many more. Classification is the fundamental problem in MIR. Our system can contribute to the automatic categorization systems where music and mood can be detected with the help of BoF representation. Music classification can help end users to select a particular song of their interest as well as on the other hand it can also help in managing different types of music more effectively and efficiently once they are categorized into different groups.

C. Overview of Audio Features

Feature extraction and audio classifier learning are the two main important tasks involved in MIR and it is the process in which a segment of an audio is represented in a compact numerical format. Audio features play a very important role in MIR because different features are associated with different classification tasks. A lot of research is still going on in extracting various audio features from the musical clip based on which we can analyze and classify a list of audio files. Different naming and describing conventions are used to classify audio features. Each convention attempts to capture audio features from certain perspective.Weihs et al.[19] said that we can classify audio features into four sub-categories namely short-term features, long-term features , semantic features and compositional features.The easiest way to divide extracted audio features into three categories as shown in figure 1,namely low-level features, mid-level features and top-level labels from the user perspective.

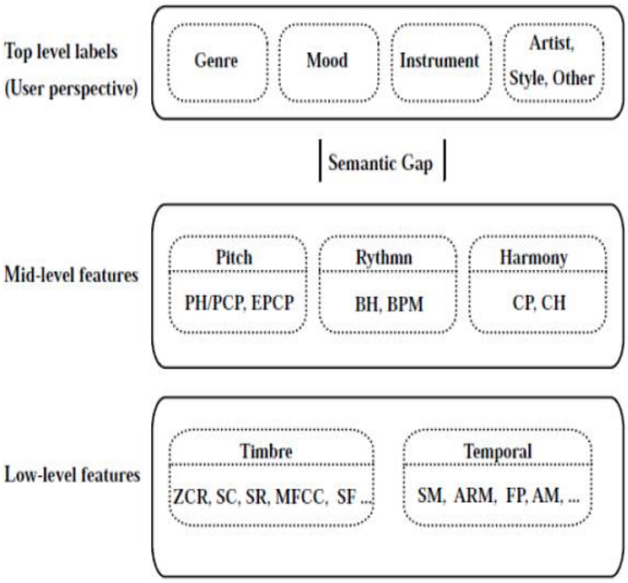

Fig.1. Audio features taxonomy

i) Low level features:- They can be further classified into two classes: Timbre and Temporal. Timbre features captures the tonal quality of a sound whereas temporal features capture the variation and evolution of timbre over time.The calculated features are based on short time fourier trasform (STFT) and are calculated for every short-time frame of sound.

ii) Mid-level features deals with pitch, rhythm and harmony which are related to the strength, tempo and regularity of a song.

iii) Top level labels gives meaning to the song from user perspective as it includes genre, mood,instrument, artist, style etc. tagging.

II. Related work

In this paper research consists of different sub-tasks namely audio feature identification and extraction, Mood model and music mining algorithms and classification.

A. Feature extraction :

According to Zhouyu Fu et al.[18] ,The key components of a classification system are feature extraction and classifier learning in which Feature extraction addresses the problem of how to represent the examples to be classified in terms of feature vector or pairwise similarities while classifier learning is to find a mapping from the feature space to the output labels so as to minimize the prediction error. They had systematically summarized the state-of-art techniques for music classification along with the recent development.

B. Music and mood classification

According to author K.C.Dewi and A.Harjoko [20],detection of mood label for the song is easy and beneficial with the help of KNN algorithm as well as self organizing maps.In [1] authors Li Su et al., includes that evaluation that compares a large number of BoF variants which helps to understand the optimal set of BoF based music classification including using logpower ,spectrogram for low-level feature representation, ODL and sparse coding for codebook learning and codeword asssignment, max pooling in the segment level,sum pooling in the clip level , logarithm tf-idf function plus entropy based idf function, and square root power normalization. This is useful for three important MIR tasks namely , genre classification, predominant instrument recognition, semantic annotation and retrieval. Chin-Chia et al[2]. stated that early temporal pooling and late dimension reduction helps to improve the efficiency of sparse coding based on audio-word extraction system. Early texture window pooling and multiple frames representation is used for early temporal pooling. M.Henaff and team[3] represented a system which extracts and learns about features from an audio clip in unsupervised manner. In this paper it is shown that Predictive sparse decomposition has ability to automatically learn useful features from constant Q spectrograms. 3 layer deep belief networks such as convolutional deep belief networks, conditional deep belief networks and convolution neural networks are used to achieve 84.3% accuracy in [4] according to P.Hamel and D.Eck. In [5] ,J Ghosh and his team compared the three clustering algorithms for Vector Quantization and found out that means achieves competitive result with relatively less computational cost.Vector quantization has been applied to genre classification. According to Aniruddha M Ujlambkar [7],11 different algorithms are used to classify music and mood on the audio file. To classify the song results 4 accuracy measures are evaluated namely ROC,F-measure,Precision and recall and it was found out that among all the algorithms used bagging approach of classification tree algorithms like RandomTree, RandomForest and SimpleCART shows improved results.

C. Mood models

Experts from musicology have came with multiple mood models which describes human emotions. According to Russell [15] a circumplex model represents human emotions on circle with each mood category plotted within the circle separated from other categories along the polar co-ordinates. Two dimensional model is suggested by Thayer[17] which is plotted based on stress and energy of the song. Henver[7] has explained categorization of moods with the help of various adjectives into 8 groups.

Proposed algorithm

In our proposed system we present music genre and mood detection system which consists of different steps. We categorize the functionality of the each step and describe interactions between individual steps.

A. *Music preprocessing*

In order to identify the genre and mood tag for an audio clip, number of audio files are provided as an input to the proposed system. In proposed system architecture .WAV and .MP3 files are used as an input because these formats are widely used now. In audio preprocessing , each input file is divided into small clips called as frames and if audio file is in .MP3 format then it is converted into .WAV format. It is important to convert .MP3file to .WAV file as it will help in assuring that files would be processed and analyzed are consistent in format and structure.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

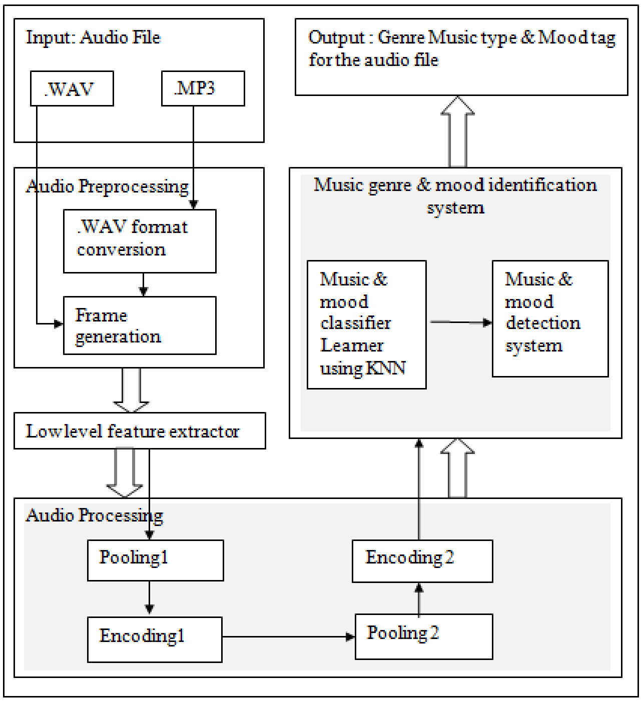
Input: Audio File
.WAV
.MP3
Output : Genre Music type & Mood tag for the audio file
Audio Preprocessing
.WAV format conversion
Frame generation
Music genre & mood identification system
Music & mood classifier Learner using KNN
Music & mood detection system
Low level feature extractor
Audio Processing
Pooling1
Encoding 2
Encoding1
Pooling 2

Fig. 2 Proposed system architecture

B. *Low level feature extractor*

Low level features such as RMS, ZCR,MFCC etc are calculated based on STFT for short span of audio clip.

C. *Audio processing*

In proposed system ,audio processing consists of temporal pooling mechanism which is defined as any functions that are able to transform the input time series representation into more temporally compact representation. Temporal pooling processing consists of following functions [2]:-

i) Early temporal pooling - When temporal pooling is applied before encoding then it is known as Early temporal pooling. Early temporal pooling refers to functions that convert raw low level representation to a more temporally compact representation. It has two important early temporal pooling functions as Early Texture Window Pooling where each segment is pooled by maximum and mean variance functions with fix 3-second window and Multiple Frames Representation where multiple consecutive frames are concatenated and used as the input vectors for feature learning algorithms.

ii) Early dimension reduction - Dimension reduction algorithm can be used to reduce the dimension of each instance.

iii) Encoding - Encoding helps in construction and interpretation of input audio signals.

iv) Late temporal pooling - When temporal pooling is applied after encoding then it is known as Late temporal pooling.

D. *Music genre and mood identification system*

After the analysis of BoF features , Music type and mood can be detected. Tagging a song with a mood label, various mood models are available. Among those mood models Thayer's mood model is best and simple to use. It describes the mood with two factors:-

i) Stress dimension (happy/anxious)

ii) Energy dimension (calm/energetic)

Thayer's mood model divides mood into four clusters namely, *Contentment, Depression, Exuberance, Anxious (Frantic).Contentment* refers to calm and happy music. *Depression* refers to calm and anxious music. *Exuberance* refers to happy and energetic music and *Anxious* refers to frantic and energetic music. Figure 3 below shows Thayer's mood model-

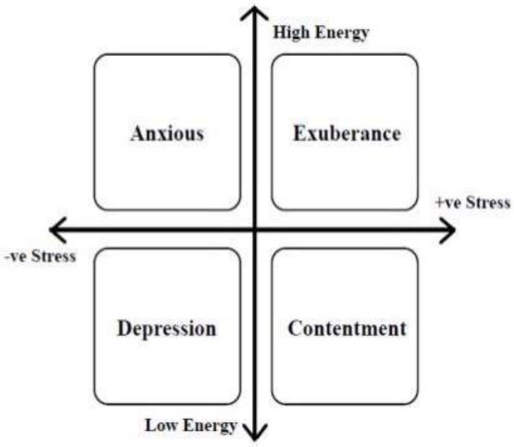

Fig. 3 Thayer's mood model

III. Mathematical model

Mathematical model for music type and mood classification system is as follows:-
Let S, be the proposed system used for categorization of song which can be represented as

$$S = \{\{I\} , \{P\} , \{O\}\}$$

Where,

1) I - Any audio file in .WAV or .MP3 format

$$I = \{X_i \mid i = 1, 2, ..., N\} \text{ where } X_i\text{= Any audio file}$$

2) P - Process of music and mood classification based on various different features. It consists of following functions:-

$$P = \{ f1 ,f2, f3, f4, f5\}$$

f1 -> Conversion of audio files into .WAV format
f2 -> Frame generation process
f3 -> Audio feature extraction
f4 -> Genre and mood classification using KNN algorithm
f5 -> Evaluation parameters (Precision, Recall, Accuracy, F-Measure)

3) O - Output as detected music type and mood label for the song

IV. Experimental setup and Results

A. Dataset used

A huge personal music collection of bollywood songs is used in .MP3 and .WAV format. Multiple open source libraries and tools are used for audio preprocessing. For training the proposed system multiple audio files are used which belongs to music type such as Classical ,Rock, Pop, Gazals , retro, Sufi etc. and mood type such as Anxious, Exuberance, Contentment and depression. Training dataset consists of 700 songs while testing dataset consists of 300 random bollywood songs.

B. Performance measure

Music and mood detection system is evaluated with the help of following evaluation measures:-

i) Confusion matrix - To evaluate the performance of the music and mood detection system widely used confusion matrix is used. It is useful for pointing the opportunities to improve the accuracy of the system. Columns of the confusion matrix represents the predictions while rows represent actual class. Accurate predictions always lie on the diagonal of the matrix. It is given by equation 1 as,

$$\begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix} \qquad \text{(Eq. 1)}$$

Here, True positives (TP) represents the number of songs that were correctly predicted , True negatives (TN) represents the number of songs not of a particular class that were correctly predicated as not belonging to that class. False Positive (FP) represents the number of songs not belonging to the class that were predicted as belonging to that class and False negatives (FN) represents the number of songs that were incorrectly predicted belonging to other class.

ii) Precision - Precision is the total percentage how many of a particular class instances as determined by proposed system exactly belong to that particular class. Precision is given by equation 2 as,

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{(Eq. 2)}$$

iii) Recall - Recall is the total percentage how many of a actual class instances as determined by proposed system. Recall is given by equation 3 as,

$$\text{Recall} = \frac{TP}{TP + FN} \qquad \text{(Eq. 3)}$$

iv) Accuracy - Accuracy is given by equation 4 as,

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad \text{(Eq. 4)}$$

iv) F-Measure - F-Measure is the harmonic mean of precision and recall. It is an average between two percentage. F-Measure is given by equation 5 as,

$$\text{F-Measure} = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}} \qquad \text{(Eq. 5)}$$

C. Results

Following table shows the utility of the above proposed system in comparison with base paper approach [1] where music type is categorised based on encoding and pooling mechanism along with lyrical processing and power normalization. Results are compared between encoding & pooling mechanism with lyrical processing and tf-idf calculation and reversing the mechanism of encoding & pooling.

	Songs correctly classified		Songs wrongly classified	
Parameters	TP	TN	FP	FN
Encoding and pooling mechanism with lyrical processing and tf-idf calculation	197	37	51	13
Encoding and pooling mechanism in reverse as per proposed approach	202	41	48	09

Table I: Confusion matrix in comparison with proposed system

Analysing the precision, recall ,accuracy and f-measure is shown in Table II, we see that using pooling and encoding mechanism and KNN classifiers , proposed system has increased the accuracy.

Parameters	Precision	Recall	F-Measure	TP-rate	FP-rate	Accuracy
Encoding and pooling mechanism with lyrical processing and tf-idf calculation	79.43%	93.80%	0.8601	83.47%	79.68%	78.6%
Encoding and pooling mechanism in reverse as per proposed approach	80.80%	95.73%	0.8763	83.12%	84.21%	81.10%

Table II: Comparative result of music and mood detection

V. Conclusion and Future Scope

In this paper music type and mood of an audio clip is detected with the help of Bag-of-frames approach where encoding and pooling mechanism has been effectively used. Any audio file can be annotated with a mood label based on Thayer's mood model comprising of four mood types : *Exuberance, Anxity, Serene and Depression.* Reversing the technique of encoding and pooling which is known as early temporal pooling and late temporal pooling helped to improve the performance of the system by approximately 2.5% in categorization of the popular Indian music. Categorization of the song using KNN algorithm has improved the accuracy of the system to great extent. We also observed that data set used to build the music and mood of the system could be increased further to improve the accuracy of the classification system as well as we can further extend the system for classification of artist, instrument etc. along with different classifiers.

References

1. Li Su,Chin-Chia Michael Yeh, Jen-Yu Liu, Ju-Chiang Wang, and Yi-Hsuan Yang,"A Systematic Evaluation of the Bag-of-Frames representation for Music Information Retrieval",IEEE ,Aug 2014.
2. Chin-Chia Michael Yeh and Yi-Hsuan Yang,"Towards a More Efficient Sparse Coding Based Audio-word Feature Extraction System",IEEE 2013.
3. M. Henaff, K. Jarrett, K. Kavukcuoglu, and Y. LeCun,"Unsupervised learning of sparse features for scalable audio classification",ISMIR, 2011.
4. P. Hamel and D. Eck,"Learning features from music audio with deep belief networks",ISMIR, 2010.
5.M. Riley, E. Heinen, and J. Ghosh,"A text retrieval approach to content-based audio retrieval",ISMIR, 2008.
6. Velankar Makarand and Dr. Sahasrabuddhe H.V.,"Novel Approach for Music search using music contents and human perception",IEEE,2014.
7. Aniruddha M Ujlambkar,"Automatic Mood Classification Model For Indian Popular Music",IEEE, 2012.
8. Vallabha Hampiholi,"A method for classification based on Perceived mood detection for Indian Bollywood Music",World Academy of Science, Engineering and Technology, Vol:72 2012-12-25.
9.Kate Hevner, (1936), "studies of the elements of expression in music", American Journal of Psychology, 48:246268.
10.Paul R. Farnsworth, (1958), "The social psychology of music, The Dryden Press.",1958.
11.Chia-Chu Liu, Yi-Hsuan Yang, Ping-Hao Wu, Homer H. Chen, (2006), "Detecting and classifying emotions in popular music", JCIS Proceedings.
12.Dalibor Mitrovic, Matthias Zeppelzauer, Horst Eidenberger, (2007), "Analysis of the Data Quality of Audio Descriptions of Environmental Sounds", Journal of Digital Information Management, 5(2):48.
13. Fu, Z., Lu, G., Ting, K. M., Zhang D., (2010), "A survey of audio-based music classification and annotation", IEEE Trans. Multimedia.
14. Geroge Tzanetakis, Perry Cook, (2002), "Musical Genre Classification of Audio Signals", IEEE Transaction on Speech and Audio Processing.
15. Russell J. A., (1980), "A circumplex model of affect",Journal of Personality and Social Psychology, 39: 1161-1178.
16. Scaringella, N., Zoia, G., Mlynek, D., (2006), "Automatic genre classification of music content, A survey.", IEEE Signal Processing Magazine, vol. 23, no.2, pp. 133141.
17. Thayer, R. E., (1989), 'The Bio-psychology of Mood and Arousal', New York: Oxford University Press.
18. Zhouyu Fu, Guojun Lu, Kai Ming Ting, Dengsheng Zhang, (2011), "A Survey of Audio-Based Music Classification and Annotation", IEEE Transactions on multimedia, Vol. 13, No. 2.
19. Weihs, C., Ligges, U., Morchen, F., Mullensiefen, D., (2007), "Classification in music research.", Advance Data Analysis Classification, vol. 1, no. 3, pp. 255291.
20. Dewi K.C., Harjoko. A., (2010), "Kid's Song Classification Based on Mood Parameters Using K-Nearest Neighbor Classification Method and Self Organizing Map.", International Conference on Distributed Frameworks for Multimedia Applications (DFmA).
21. IEEE online,http://ieeexplore.ieee.org
22. ISMIR online, http://www.ismir.net/

BIOGRAPHY

HARSHALI NEMADE received the bachelors of computer engineering degree from All India Shri Shivaji Memorial Society's Institute of Information technology in 2011 from Pune University,Pune. She is now pursuing Masters of engineering degree in computer engineering from P.E.S's Modern college of engineering, Savitribai Phule Pune Univerity,Pune.

DOI: 10.15680/IJIRCCE.2017. 0506225