Connotative Feature Extraction For Movie Recommendation

N. G. Meshram *meshram.naina1@gmail.com*
PG Department of Computer Science and Engineering
Prof Ram Meghe College of Engineering and Management
Badnera, 444 701, India

A. P. Bhagat *amol.bhagat84@gmail.com*
PG Department of Computer Science and Engineering
Prof Ram Meghe College of Engineering and Management
Badnera, 444 701, India

Abstract

It is difficult to assess the emotions subject to the emotional responses to the content of the film by exploring the film connotative properties. Connotation is used to represent the emotions described by the audiovisual descriptors so that it predicts the emotional reaction of user. The connotative features can be used for the recommendation of movies. There are various methodologies for the recommendation of movies. This paper gives comparative analysis of some of these methods. This paper introduces some of the audio features that can be useful in the analysis of the emotions represented in the movie scenes. The video features can be mapped with emotions. This paper provides methodology for mapping audio features with some emotional states such as happiness, sleepiness, excitement, sadness, relaxation, anger, distress, fear, tension, boredom, comedy and fight. In this paper movie's audio is used for connotative feature extraction which is extended to recognize emotions. This paper also provides comparative analysis of some of the methods that can be used for the recommendation of movies based on user's emotions.

Keywords: Audio Features, Connotative Features, Emotion Recognition, Movie Recommendation, Video Features.

1. INTRODUCTION

Due to advancement in multimedia devices, it is easy to access the private content of multimedia such as how the people enjoy the movies. Decision of which movie to watch is most of the time taken from friends suggestions. Nowadays, one can take the benefit of media recommender system [1, 2] which has the ability to suggest the video content on the basis of person's current affective state, social experiences and profile. The psychologist has investigated on the emotional properties of the film media in terms of empathy with situation and characters and also in terms of director's establishment of film making techniques which provides the emotional cues. The empathy is not with the characters which provide the affective cues with film media while film makers make the use of techniques such as editing, musical scores, lighting so as to emphasize the particular emotional interpretation by the viewer [3]. This is referred as connotation which gives the path of communication and influences how the meaning is transmitted to the audience which is conveyed by the director.

An expert system MovieGEN [4] is used for the recommendation of movies. The user's information is taken as input and their movie preferences are predicted using the support vector machine and on the basis of prediction movies are selected from the dataset and generate some questions to the users. On the user's answer it the movie is recommended for the user. Recommendation systems are the expert system where the knowledge of expert is combined with user's preferences to filter the information and provide the users with the information. There

are two main approaches for filtering collaborative and content based approach. Most of the recommendation systems use the hybrid approach combination of these two approaches. The model using SVM based learning techniques [4, 5] is used in MovieGEN. With the help of this model one can predict genres and period of movies that user prefers. The implementation of content based approach is provided in [4, 5] such that it takes into consideration the user choice which is not based on the user's past history but on the answers the user gives to the question. Collaborative approach, knowledge based approach, hybrid approach, etc. based recommendation systems have been developed for different types of domains [6]. In online movie recommendation system MovieLens [7] the first time when user logins, the system ask to rate certain movies to the user which the user has seen and these ratings are recommended with the other movies that the user has not seen. This type of filtering is based on rating which uses collaborative approach.

Recommendation system separates the relevant content from the non-relevant content which is based on the individual user's preferences are presented in [8]. The content items are described with the metadata in the content based recommender system and are stored in the item profile. There are two approaches for affective labeling that is explicit and implicit. The explicit approach provides the unambiguous labels and the drawback of this approach is the intrusiveness of the process. The implicit affective labeling is unobtrusive and it is not affected by the user's personal motives. Movies were used as the core label for building the user preferences and recommending the movie user-centric approach for labeling the content and building the user profile were used in recommender system. A framework for affective labels in recommender system where three stages were distinguish in interaction with the multimedia content such as context, induced emotion and implicit rating is proposed in [9].

In implicit labeling the most popular approach is to expose the user's to stimuli and record their responses [10, 11]. By using the emotion detection techniques the affective labels can be detected. Such that the [12] used the 2-minute video clips to expose the user's and recorded the different types of physiological signals and also the ground truth emotive response in valence-arousal plane explicitly. The features which are used commonly are the geometric features like active appearance model features and facial points and the various classification techniques includes support vector machine [12, 13]. The [14] took the advantage of facial expression, audio features, and shoulder tracking to predict the affect in valence arousal space. However, there is a need to model the response of complex approach which is the future challenge.

According to [15] emotions are personal and everyone reacts to the events or the media content that depends on cultural, personal, short term and subjective factors. For example, when two people watches the same horror movie and reacts differently or they may share similar view on movie from their individual affective response. Connotations are linked to the emotions. Emerging theories of filmic emotions elicit the mechanism that inform mapping between video features and emotional models. Filmic emotions are less character oriented giving a greater prominence to style and argues that moods generate the expectations of particular emotional cues and concluded that emotional associations provided by music encourage to relate the video features to emotional responses is suggested in [16]. The way to assess the affective dimension of media is by the use of expected mood is proposed in [17]. Film maker tries to communicate the set of emotions when they produce the movie for the audience. The emotional clustering of films for different genres is described in [16, 17]. This clustering approach may be used to target the user emotions. The audio and visual low level features in a high dimensional space to extract the meaningful patterns by SVM inference engine is proposed in [15]. It states audio cues are more informative than visual with respect to the affective content.

During human to human interaction changes in the person's affective state plays an important role. In affective computing one of the applications can be human computer interaction. The learning system called "Multimodal Human Computer Interaction: Towards a Proactive Computer" is discussed in [13]. In this type of learning environment the user is able to explore the games by interacting with the computer avatar. In this environment multiple sensors were used to detect

and track the behavioral cues of user, camera was used to record the facial expression of user, to track the eye movements , to monitor the task progress and a microphone was used to record the speech signals. And based on this over all information, the avatar offers an appropriate strategy in this type of learning environment. The psychological studies [1], [16] indicated that while judging someone's affective state, people rely on the facial expression and the vocal intonations. The motivation for the audio-visual fusion is the improved reliability. Current techniques which are used for the detection and tracking of facial expression are very sensitive to the clutter, head pose and variations in lighting condition and the current techniques which are used for the speech processing are very sensitive to auditory noise. The table 1 shows the comparative analysis of available movie recommendation methodologies.

Title	Methodology	Feature Extraction	Data Used	Experiment	System	Measures
Affective Recommendation of Movies Based on Selected Connotative Features	Scene ratings by users Feature extraction Feature selection Regression Scene recommendation	Visual, audio, film grammar, color	Movie scenes	SVR	SVR model	Connotative space, precision For example, precision for Top -3 close and far scenes at different emotional distance is given by (Top-3 Close) d=0, (precision@3) 0.30, (Top-3 Far) d=4, (precision@3) 0.22.
MovieGEN: A Movie Recommendation System	Machine learning based preference prediction	Vector format (sample is quantified into vector of integer number)	Movie preferences	SVM Correlation analysis using SVM regression	Machine learning model	Training data, preference vector, clustering
Affective Labeling in a Content-Based Recommender System for Images	Emotion detection evaluation methodology Affective CBR system evaluation methodology	Active appearance model feature, facial feature, audio features	images	SVM, NaiveBayes, AdaBoost,C4.5	CBR system	Accuracy, confusion matrix, classifier For example , for explicit affective labeling classifier SVM (Precision) 0.68, (Recall) 0.54, (F)measure 0.60
A Connotative Space for Supporting Movie Affective Recommendation	Span the semantic space Validate by inter-rater agreement Support affective recommendation	Video features, emotional responses, audio features, visual features	Movie segments, video frames movie scenes	Osgood's semantic space	SVM inference engine	Connotative space, emotional wheel
Audio-Visual Affective Expression Recognition Through Multistream Fused HMM	MFHMM HCI	Facial features, audio features, multistream fused HMM	images	Face only HMM, pitch only HMM, energy only HMM, independent only HMM, MFHMM	Motion units	Expression, performance For example, MFHMM for (happy) 0.70
Affective Video Content Representation and Modeling	Affective level Cognitive level	Movie facts, feelings or emotions	Movie scenes, video clips	Video content	Arousal, valence	Arousal and valence
Robust Face-Name Graph Matching for Movie Character Identification	Character identification Face-name Graph matching	To identify faces of characters	movie	Face-Name graph matching	Error correcting graph matching	Face track Detection Accuracy clip 1 (# Face track) 372, (# Track detected) 352, (Accuracy) 95.2%
SMERS: Music Emotion Recognition Using Support Vector Regression	Feature extraction Mapping	Pitch, tempo, loudness, tonality, key, rhythm and harmonics	Music and emotion	SVM,SVR and GMM	Thayer's two-dimensional emotion	SVR training parameters and obtained optimums in polar representation Name of parameters:

	Training				model	Distance Angle Nu (u) : 2^{-8} , 2^{-8} ,Gamma of RBF (g) : 2^{-10} , 2^{-4} Cost (C) : 2^{8} , 2^{6} mean squared error: 0.02498 ,0.09834

TABLE 1: Comparative analysis of available movie recommendation methods.

The presence of entrainment at the emotion level in cross-modality settings and its implications on multimodal emotion recognition systems is investigated in [18]. The relationship between acoustic features of the speaker and facial expressions of the interlocutor during dyadic interactions are explored. More than 72 % speakers displayed similar emotions, indicating strong mutual influence in their expressive behaviors. The cross-modality, cross-speaker dependency, using mutual information framework is also investigated. A strong relation between facial and acoustic features of one subject with the emotional state of the other subject is revealed. It has been suggested that the expressive behaviors from one dialog partner provide complementary information to recognize the emotional state of the other dialog partner. Classification experiments exploited cross-modality, cross-speaker information. The emotion recognition experimentations are demonstrated using the IEMOCAP [19] and SEMAINE [20] databases.

In this paper the audio features CEP and SPEC are used to extract the emotions from the audiovisual datasets. These extracted features are used to affection based movie recommendations. An experimental result shows that the proposed methodology can be utilized to provide precise and effective recommendation results in faster manner. The total time required for recommendation can be reduced due to use of CEP and SPEC features.

2. CONNOTATIVE FEATURE EXTRACTION

The audio features such as sound, voice and music play an important role of expression in shaping the scene affection of the audience. The algorithm called as Piecewise Bezier Volume Deformation (PBVD) [5] tracking can be use to extract the facial features. It uses the 3-D facial mesh model that is embedded in multiple Bezier volumes. With the help of the movements of control points the shape of mesh can be changed in Bezier volumes which guarantees that the surface patch to be smooth and continuous.

Entropic Signal Processing system is used as a software package for the audio feature extraction. It implements the algorithm using cross correlation function and dynamic programming. In the experimental results [3], [5] it is suggested that pitch and energy are the important factors in affect classification. The pitch varies from person to person. Males speak with lower pitch than the females.

Multistream Fused Harmonic Markov Model can be used for integrating audio and visual feature which is used to construct new structure for linking the multiple components HMM according to maximum entropy principle and maximum mutual information. It is the generalization of two-stream fused HMM. Such that MFHMM is used for recognition problem with more than the two feature streams.

The problem of emotions which is connected to the use of other people's affective annotations while the connotative properties agreed by the people's emotional reaction provides accurate recommendation methods and to establish the method for performing research on emotions such as users behavior, users self reporting and learning method shows how to translate the low level and mid-level properties of videos into inter-subjective for affective analysis of film. The emotional characters of videos are used to study the narrow set of situations such as for the sporting events or the horror movies. The advantage of ranking is based on similarities between items which are close to human emotions instead of using absolute labeling.

3. AUDIO ANALYSIS USING CEPSTRUM (CEP) AND SPECTRUM (SPEC)

A Cepstrum (Cep) is the result of taking the Inverse Fourier transform (IFT) of the logarithm of the estimated spectrum of a signal. There is a complex cepstrum, areal cepstrum, a power cepstrum, and phase cepstrum. The power cepstrum in particular finds applications in the analysis of human speech. The name "cepstrum" was derived by reversing the first four letters of "spectrum". Operations on cepstra are labelled quefrency analysis, liftering, or cepstral analysis.

The power cepstrum of a signal is defined as the squared magnitude of the inverse Fourier transform of the logarithm of the squared magnitude of the Fourier transform of a signal. Mathematically represented as

$$| F^{-1}\{\log(|F\{f(t)\}|^2)\}|^2$$

The complex cepstrum holds information about magnitude and phase of the initial spectrum, allowing the reconstruction of the signal. The real cepstrum uses only the information of the magnitude of the spectrum. The process is defined as FT → abs() → log → IFT, i.e., that the cepstrum is the "inverse Fourier transform of the log-magnitude Fourier spectrum". The cepstrum is a representation used in homomorphic signal processing, to convert signals (such as a source and filter) combined by convolution into sums of their cepstra, for linear separation. In particular, the power cepstrum is often used as a feature vector for representing the human voice and musical signals. For these applications, the spectrum is usually first transformed using the mel scale. The result is called the mel-frequency cepstrum or MFC (its coefficients are called mel-frequency cepstral coefficients, or MFCCs). It is used for voice identification, pitch detection and much more. The cepstrum is useful in these applications because the low-frequency periodic excitation from the vocal cords and the formant filtering of the vocal tract, which convolve in the time domain and multiply in the frequency domain, are additive and in different regions in the quefrency domain.

The frequency spectrum of a time-domain signal is a representation of that signal in the frequency domain. The frequency spectrum can be generated via a Fourier transform of the signal, and the resulting values are usually presented as amplitude and phase, both plotted versus frequency. A source of sound can have many different frequencies mixed. A musical tone's timbre is characterized by its harmonic spectrum. Sound spectrum is one of the determinants of the timbre or quality of a sound or note.

Spectrum (Spec) analysis, also referred to as frequency domain analysis or spectral density estimation, is the technical process of decomposing a complex signal into simpler parts. As described above, many physical processes are best described as a sum of many individual frequency components. Any process that quantifies the various amounts (e.g. amplitudes, powers, intensities, or phases), versus frequency can be called spectrum analysis. Spectrum analysis can be performed on the entire signal. Alternatively, a signal can be broken into short segments (sometimes called frames), and spectrum analysis may be applied to these individual segments. Periodic functions (such as sin(t)) are particularly well-suited for this sub-division. General mathematical techniques for analyzing non-periodic functions fall into the category of Fourier analysis. The Fourier transform of a function produces a frequency spectrum which contains all of the information about the original signal, but in a different form. This means that the original function can be completely reconstructed (synthesized) by aninverse Fourier transform. For perfect reconstruction, the spectrum analyzer must preserve both the amplitude and phase of each frequency component. These two pieces of information can be represented as a 2-dimensional vector, as a complex number, or as magnitude (amplitude) and phase in polar coordinates. A common technique in signal processing is to consider the squared amplitude, or power; in this case the resulting plot is referred to as a power spectrum.

In practice, nearly all software and electronic devices that generate frequency spectra apply a fast Fourier transform (FFT), which is a specific mathematical approximation to the full integral solution. Formally stated, the FFT is a method for computing the discrete Fourier transform of

a sampled signal. Because of reversibility, the Fourier transform is called a representation of the function, in terms of frequency instead of time; thus, it is a frequency domain representation. Linear operations that could be performed in the time domain have counterparts that can often be performed more easily in the frequency domain. Frequency analysis also simplifies the understanding and interpretation of the effects of various time-domain operations, both linear and non-linear. Some kind of averaging is required in order to create a clear picture of the underlying frequency content (frequency distribution). Typically, the data is divided into time-segments of a chosen duration, and transforms are performed on each one. Then the magnitude or (usually) squared-magnitude components of the transforms are summed into an average transform. This is a very common operation performed on digitally sampled time-domain data, using the discrete Fourier transform. Such processing techniques often reveal spectral content even among data which appears noisy in the time domain.

4. PROPOSED CONNOTATIVE FEATURE EXTRACTIONS FOR MOVIE RECOMMENDATION

Figure 1 shows the overall system flow which includes the extraction of audio features. There are different types of movies such as Horror, Action, Thriller, Comedy, Animation etc. So, all these different types of movies or movie scenes will be displayed to the multiple users. From these movies/movie scenes the audio will be extracted. That is, to extract the connotative features from the extracted audio using support vector regression (SVR). These extracted audio features will be mapped with the connotative features. SVR will rate/rank according to the connotative features. Then SVR will ask or query to the users. After asking questions SVR will compare the result with the actual user preferences. Then the user will modify the rate/rank of the connotative features. At last, SVR will recommend the movie to the user.

One of the first decisions in any pattern recognition system is the choice of what features to use: How exactly to represent the basic signal that is to be classified, in order to make the classification algorithm's job easiest. Speech recognition is a typical example. The most popular feature representation currently used is the Mel-frequency Cepstral Coefficients or MFCC. Another popular speech feature representation is known as RASTA-PLP, an acronym for Relative Spectral Transform - Perceptual Linear Prediction. PLP was originally proposed by Hynek Hermansky [10] as a way of warping spectra to minimize the differences between speakers while preserving the important speech information. RASTA is a separate technique that applies a band-pass filter to the energy in each frequency sub-band in order to smooth over short-term noise variations and to remove any constant offset resulting from static spectral coloration in the speech channel e.g. from a telephone line. Machine learning focuses on the prediction based on the known properties learned from the training data and utilized in data mining and knowledge discovery domains. The data sets for any machine learning model are parts such as input and output. The support vector machine based on statistical learning can be used to overcome the problems such as over-fitting, local minimum and sufficient for high generalization. The introduction of fuzzy logic into SVM created a new multi-layer SVM which can be applied later into the numerical regression problem.

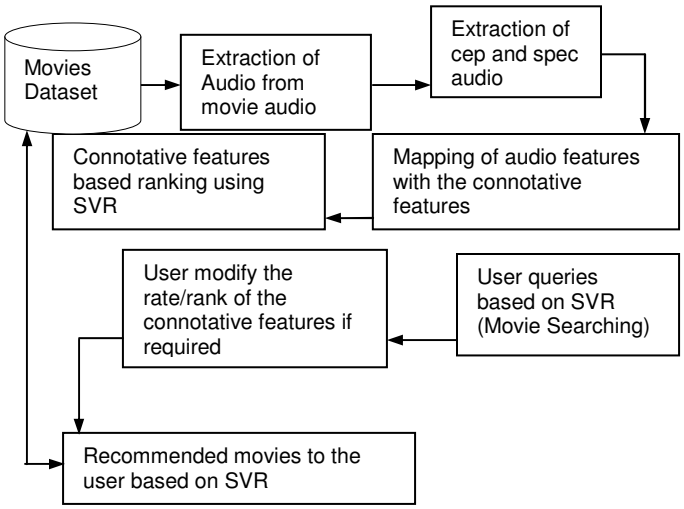

FIGURE 1: Proposed Connotative Feature Based Movie Recommendation System Model.

The proposed approach for recommendation of movies works as explained here. Initially for the extraction audio features from the move scenes the audio is extracted from the movies. After the extraction of audio the cep and spec features are extracted from the audio. The set of movies which belong to the popular films is considered and asked users to rate each of these scenes on three connotative dimensions. The distances are computed using the Earth mover's distance on rate histogram axis (N, T, E) as:

$$\Delta_{i,j}^{x} = EMD(\mathrm{H}_i^x, \mathrm{H}_j^x)\ \mathrm{x} \in \{\mathrm{N}, \mathrm{T}, \mathrm{E}\}$$

Support vector regression is used to relate connotative distances on the user's rates. The connotative distances are predicted between the movie scenes when the model is validated. With the help of SVR model the connotative distances are predicted .The user has to choose the query item and the scenes which have small connotative distance from query are then recommended to the users.

5. EXPERIMENTAL RESULTS

The connotative features happiness, sleepiness, excitement, fight, etc are mapped with the extracted audio features from the movie scenes. Some of the sample extracted audio features and their cep and spec values are presented below. The extracted happiness audio feature is shown in figure 2 and its cep and spec values are shown in table 2. The extracted sleepiness audio feature is shown in figure 3 and its cep and spec values are shown in table 3. The extracted excitement audio feature is shown in figure 4 and its cep and spec values are shown in table 4. The extracted fight audio feature is shown in figure 5 and its cep and spec values are shown in table 5.

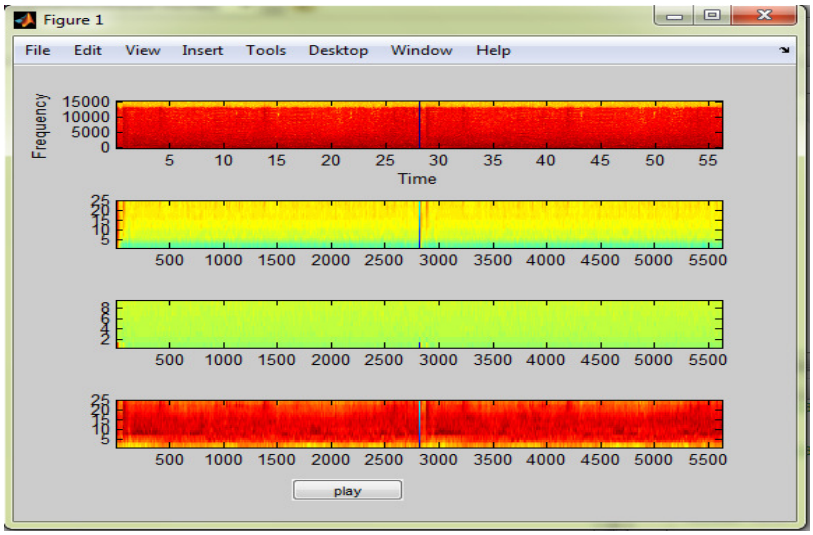

FIGURE 2: Snapshot of Happiness Audio Features Extracted From Movie Scenes.

Audio Feature	Min Value	Max Value
Cep1	-5.7452	4.2331
Cep2	-0.8349	8.6520
Spec1	3.031	10.03
Spec2	1.2945	1.609

TABLE 2: Cep and Spec Features Extracted for Happiness Movie Scenes.

After getting the audio feature values these values are mapped with the connotative features as shown in the table 6. For getting the exact mapping of the audio features with the connotative features 25 videos of each category i.e. happy, sleepy, excitement, sad, etc. are used for extracting audio features. Then by identifying the similarity between the extracted features as per movie category it has been mapped with the connotative features.

6. CONCLUSION AND FUTURE SCOPE

Feature selection algorithms are popular in different disciplines such as array analysis and multimedia analysis. The main advantage is the reduction of number of features processed and better understanding of problem. Automatic emotion recognition from the movie scenes by using the audio features by using different machine learning classification algorithms such as SVM, SVR and HMM are analyzed in this paper. SVR is the best option for the selection of the connotative feature selection and mapping for the recommendation of movie scenes. From the experimental results it has been cleared that the audio features can be successfully utilized for the affection based recommendation of movies.

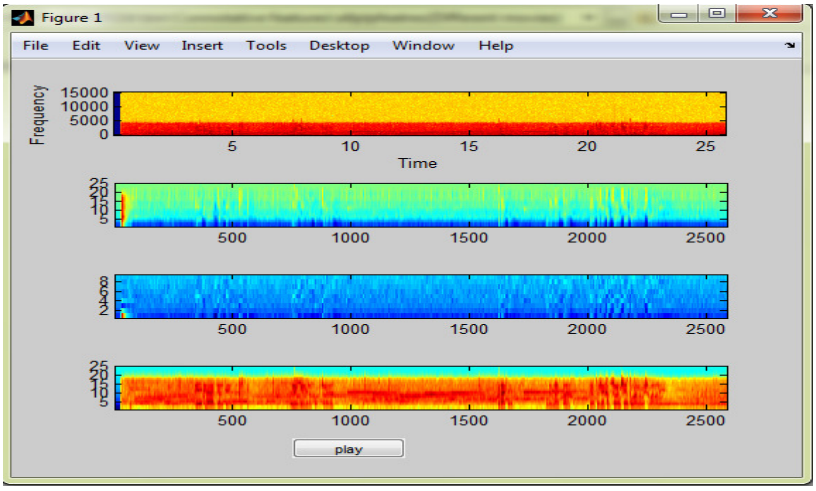

FIGURE 3: Snapshot of Sleepiness Audio Features Extracted From Movie Scenes.

Audio Feature	Min Value	Max Value
Cep1	-1.1378	2.3878
Cep2	-1.4271	7.3225
Spec1	0.0305	35.1272
Spec2	1.2945	1.1354

TABLE 3: Cep and Spec Features Extracted for Sleepiness Movie Scenes.

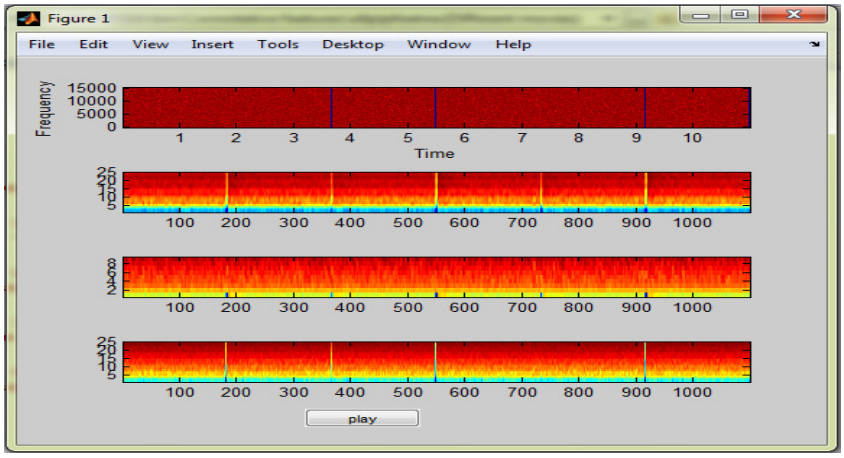

FIGURE 4: Snapshot of Excitement Audio Features Extracted From Movie Scenes.

Audio Feature	Min Value	Max Value
Cep1	-1.3713	0.0692
Cep2	-0.7832	8.9013
Spec1	0.0409	1.2208
Spec2	54.6322	2.3649

TABLE 4: Cep and Spec Features Extracted for Excitement Movie Scenes.

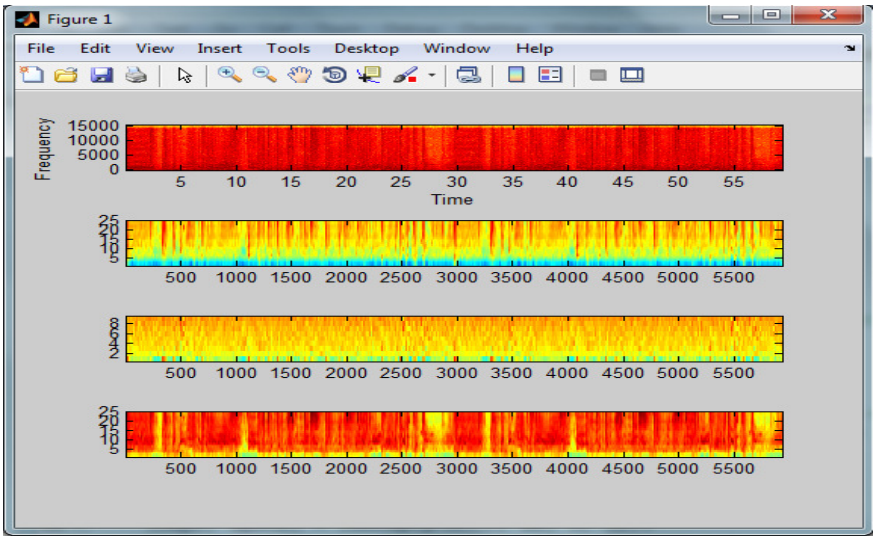

FIGURE 5: Snapshot of Fight Audio Features Extracted From Movie Scenes.

Audio Feature	Min Value	Max Value
Cep1	-2.3014	0.8861
Cep2	-0.8789	8.1108
Spec1	0.0144	6.4190
Spec2	1.3094	1.0660

TABLE 5: Cep and Spec Features Extracted for Fight Movie Scenes.

Connotative Features	Audio Features			
	Cep1	Spec1	Cep2	Spec2
Happiness	-0.75605	51.7108	3.90855	1.45185
Sleepiness	0.625	17.57885	2.9477	1.21495
Excitement	-0.65105	0.63085	4.05905	28.49855
Sadness	-1.10775	6.641	3.2043	1.5155
Relaxation	-0.5584	13.91215	3.47175	1.25855
Anger	-0.0658	11.6543	2.59565	1.92725
Distress	-0.64005	6.40145	3.1438	1.25915
Fear	-0.7565	11.5238	3.20855	1.1733
Tension	-0.77315	81.0031	3.5785	1.4281
Boredom	-0.7664	9.45755	2.85995	3.2319
Comedy	-1.1063	16.264	3.5623	1.3652
Fight	-0.70765	3.2167	3.61595	1.1877

TABLE 6: Mapping of Audio Features with Connotative Features.

Other classification algorithms such as fuzzy and KNN can be considered for the further research. The future plans are to compare the results of machine learning based emotion recognition with human performed arousal and valence data. This paper compares various strategies for the connotative feature based movie recommendations. This initial comparison is helpful for the researchers to get basics of affective recommendations of movies. This paper also presents how the audio features can be mapped with the connotative features so that these features can be

utilized for the emotion recognition. The proposed strategy can be extended so that it can support the web based interface to provide the relevant movies as per user query.

7. REFERENCES

[1] G. M. Smith, "Film Structure and the Emotion System". *Cambridge, U.K.: Cambridge Univ. Press, 2003*.

[2] E. A. Eyjolfsdottir, G. Tilak, N. Li, "MovieGEN: A Movie Rec System," IEEE Trans on Mult., Vol 15, no 5, Aug 2010.

[3] J. Kim and E. Andre," Emotion Recognition Based on Physiological Changes in Music Listening," IEEE Trans on Pattern Analysis and Machine Intelligence,Vol. 30, no. 12, Dec 2008.

[4] L. Canini, S. Benini, and R. Leonardi, "Affective Recommendation of Movies Based on Selected Connotative Features," IEEE Trans on circuits and systems for video technology, vol. 23, no. 4, April 2013.

[5] A. Tawari and M. Trivedi," Face Expression Recognition by Cross Modal Data Association," IEEE Trans on Multimedia, Vol. 15, no. 7, Nov 2013.

[6] A.Smola and B. Sch¨olkopf,"A Tutorial on Support Vector Regression", Sep 2003.

[7] L. Lu, D. Liu and H. Zhang,"Automatic Mood Detection and Tracking of Music Audio Signals," IEEE Trans on audio, speech and language processing, Vol. 14, no. 1, Jan 2006.

[8] A. Metallinou, A. Katsamanis, F. Eyben and S. Narayanan," Context-Sensitive Learning for Enhanced Audiovisual Emotion Classification" IEEE Trans on affective computing, Vol. 3, no. 2, Apr-Jun 2012.

[9] Z. Deng, U. Neumann, T. Kim, M. Bulut, and S. Narayanan, "Expressive Facial Animation Synthesis by Learning Speech Coarticulation and Expression Spaces," IEEE Trans on visualization and computer graphics Vol. 12, No. 6 Dec 2006.

[10] M. Aeberhard, S. Schlichthärle, N. Kaempchen, and T. Bertram," Track-to-Track Fusion with Asynchronous Sensors Using Information Matrix Fusion for Surround to-Track Fusion with Asynchronous Sensors Using Information Matrix Fusion for Surround Environment Perception," IEEE Trans on intelligent transportation system, Vol. 13, no. 4, Dec 2012.

[11] W. Xu, C. Chang,Y. S. Hung and P. Fung," Asymptotic Properties of Order Statistics Correlation Coefficient in the Normal Cases," IEEE Trans on signal processing, Vol. 56, no. 6, Jun 2008.

[12] C. Tsai, L. Kang, C. Lin and W. Lin," Scene-Based Movie Summarization via Role-Community Networks," IEEE Trans On Circuits and Systems for Video Technology, 2013.

[13] X. Zhang, W. Hu, H. Bao, and S. Maybank, " Robust Head Tracking Based on Multiple Cues Fusion in the Kernel- Bayesian Framework" IEEE Trans On Circuits and Systems for Video Technology, Vol. 23, No. 7, July 2013.

[14] J. Kim and E. Andre," Emotion Recognition Based on Physiological Changes in Music Listening" IEEE Trans On Pattern Analysis and Machine Intelligence , Vol. 30, No. 12, Dec 2008.

[15] A. Hanjalic and L. Xu, "Affective video content representation and modeling," IEEE Trans. Multimedia, vol. 7, no. 1, pp. 143–154, Feb. 2005.

[16] C. Liu, D. Wang, J. Zhu, and B. Zhang, "Learning a Contextual Multi-Thread Model for Movie/TV Scene Segmentation," IEEE Trans on Multimedia, Vol 15, no.4, June 2013.

[17] D. Lottridge, M. Chignell, and M. Yasumura, "Identifying Emotion through Implicit and Explicit Measures: Cultural Differences, Cognitive Load, and Immersion," IEEE Transaction on Affective Computing, Vol 3, no. 2, April-June 2012.

[18] Soroosh Mariooryad and Carlos Busso," Exploring Cross-Modality Affective Reactions for Audiovisual Emotion Recognition," IEEE Transaction on affective computing, Vol. 4, no. 2, April-June 2013.

[19] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Journal of Language Resources and Evaluation, vol. 42, no. 4, pp. 335–359, December 2008.

[20] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schr¨oder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 5–17, January-March 2012.