Automatic Mood Detection of Indian Music Using MFCCs and K-means Algorithm

Garima Vyas, Malay Kishore Dutta
Dept.of Electronics and Communication Engineering
Amity University, Uttar Pradesh , Noida -201303, India
gvyas@amity.edu, mkdutta@amity.edu

Abstract- **This paper proposes a method of identifying the mood underlying a piece of music by extracting suitable and robust features from music clip. To recognize the mood, K-means clustering and global thresholding was used. Three features were amalgamated to decide the mood tag of the musical piece. Mel frequency cepstral coefficients, frame energy and peak difference are the features of interest. These features were used for clustering and further achieving silhouette plot which formed the basis of deciding the limits of threshold for classification. Experiments were performed on a database of audio clips of various categories. The accuracy of the mood extracted is around 90% indicating that the proposed technique provides encouraging results.**

Keywords- ***Mood Detection; Mel Frequency Cepstral Coefficients; Frame Energy; Peak Detection; clustering; silhouette plot.***

I. INTRODUCTION

Music has been has been an inherent part of recreation of human life. In the present scenario, available collection of music may figure up to massive number of records worldwide which pursue to mushroom each day. With so much of variety of music readily available, we humans do not always listen to the same type of music all the time. The choice is very much genuinely regulated by person's emotional state at that peculiar instant. Thus, an additional parameter or rather search filter - Mood – which signifies the emotion of that particular music piece is required. However, classifying music as per its mood is comparatively a harder task as musical mood is quite subjective term. This paper propose a method to design a system that efficiently classifies the audio signals on the basis of the emotional content i.e. mood. Feature extraction forms the basis to classify the mood. Thresholding was found to be a good method to carry out the work. The decision regarding limits was supported by clustering and further by silhouette plot.

To study the relation of music with mood, various mood models were studied like the Henver model [1], Russel model [2] etc. One method suggested the derivation of bass-line and rhythm patterns, where the analyses of the specified unit patterns are utilized for the analytical deduction of features for mood classification [3]. Another approach involved “jAudio” which is an open-source audio feature extraction framework [7]. Mood parameter based classification using Support Vector Regression [6], k-means clustering [2], Support Vector Machine [8], [9], [10] etc. were analyzed.

The main contribution of this paper is a method to accurately attach mood tag to a musical audio. High accuracy is achieved by amalgamating three features: Mel frequency cepstral coefficients (MFCCs), frame energy and peak difference together. The complexity of the technique is less as compared to the existing methods. Moreover, the scope of the proposed method is high as it reduces the manual overhead and the complexity to find an audio of particular emotion from a huge set of data. Thus, it would save not only time but also human efforts. This work has the scope to be extended to generate a playlist automatically based on mood of the music. The proposed scheme overcomes the constraints faced by existing algorithms as it uses more than one audio feature and mood is detected by decision fusion scheme.

The rest of the paper is organized as follows. Section II describes the methodology followed by the design. Section III describes the audio features used in this paper. The next section describes the clustering. Section V explains the decision rule for detecting mood. Section VI displays the simulation of results. Finally, the last section draws the conclusion and discusses the future work which may be possible in this area of work.

II. DESCRIPTION

This area of research can be subdivided into three basic and different sub-fields:

1. Mood model [1], [3], [8], [11] which demands recognizing and defining the list of adjectives precisely describing all possible moods.
2. Identification of audio features [5] and its extraction, which involves discovering and extracting the suitable features from an audio clip.
3. Data Mining [7] and related algorithm, which involves learning and selecting the appropriate algorithm(s) that help to mine effectively the music datasets with substantial accuracy.

To achieve the above mentioned goals the methodology to be followed is explained by following flow chart shown in Fig. 1.

A. Audio Pre-processor

978-1-4799-5173-4/14/$31.00 ©2014 IEEE

This component performs the task of pre-processing the audio clips provided by the user to the system. The pre-processing job involves following steps:-

Step 1: Audio File Splitting: An audio with duration of 45 sec is clipped. Such a clip has proved to be quite good from experimentation point of view as it is not very short to lose any important information and not very long to increase the processing time.

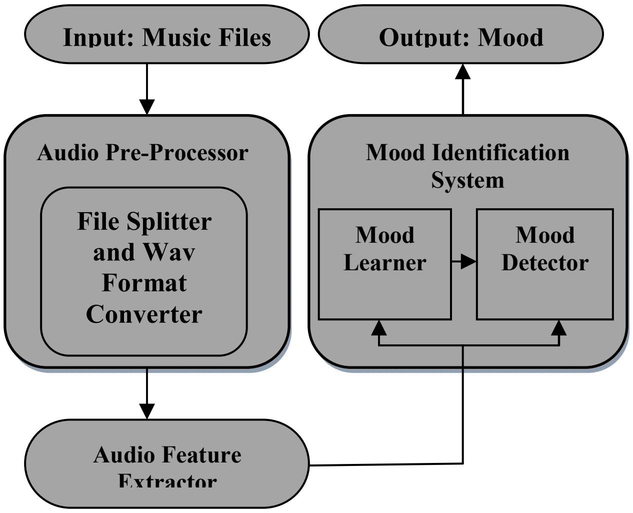


Fig. 1: Mood Detection System

Step 2: Audio Format Conversion: Each music clip of 45 second is transformed to a standard format namely WAV (type as stereo and 16 bits PCM) with a rate of sampling 44.1 kHz. Thus, this component makes sure that the input music clips given by the user are transformed so that they can be prepared for processing and analyzing.

B. Audio Feature Extractor

This specific module revolves around the features of audio signal associated with the music clips obtained as a result from the Audio Pre-processor. The audio clip of 45sec is segmented into frames of 55~60ms with an overlap of 35~40ms. From each frame the MFCCs (Mel frequency cepstral coefficients) and energy is computed. Then the maximum and minimum peak is calculated which in turns gives the peak difference.

C. Mood Identification System

This is the major processing unit of the whole system and is accountable for mining the mood from the music dataset acquired as input from the audio feature extractor module. The module has two important functions to perform as mentioned below:-

1. Mood Learner: Here, the input is received in the form of a training dataset of music features with the manually updated 'mood' attribute by the domain experts, from the training point of view. In this case clustering is used to train the system.
2. Mood Detector: The Mood detector is used to evaluate the dataset under consideration against the mood classifier model that is finalized. The evaluation uses the threshold value to classify the mood of an audio. In case a full song is fed instead by the user, the system reciprocates the maximum voted mood from the moods predicted for the clips derived from that song. The final output of this module is generally utilized by the end-user application like any Music information retrieval application or a mood-annotator.

III. AUDIO FEATURES

1. *MFCCs:*

The frames with duration of 55~60 ms is now passed through a window. Windowing minimizes the distortion in spectra and making the signal equivalent to zero at the start as well as end of every frame leading in reduction of the signal's discontinuities. Most often, the Hamming window is utilized for the purpose, defined in Eq. 1 as follows:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \qquad (1)$$

Now, on the windowed signal the Discrete Fourier Transform (DFT) is performed to analyze the different timbers in terms of frequency by using the given equation:

$$X(k) = \sum_{n=0}^{N-1} X(n) e^{-\frac{j2\Pi kn}{N}}, \quad 0 \le k \le N-1 \qquad (2)$$

The basis of MFCCs is the known and accepted fact that variation in critical bandwidth of the human ear with regards to frequency where filters are placed in a linear fashion at frequencies that are low while they are placed logarithmically where frequencies are high. They are conveyed through the scale of mel frequency. The Mel conversion takes place as written in Eq. 3:

$$Mel(f) = 2595\ log_{10}\left(1 + \frac{f}{700}\right) \qquad (3)$$

The frequencies are passed through a sequence of bandpass filters that are triangular in nature whose basis is the Mel scale. These filtered frequencies are converted back into a time-like domain called quefrency by applying discrete cosine transform (DCT) on it. DCT is chosen because it has excellent energy compaction as maximum of the energies are concentrated on low frequency components. The Eq. 4 describing DCT is:

$$Y(k) = \sum_{n=0}^{N-1} \cos[\frac{\Pi}{n}(n + \frac{1}{2})k] \quad 0 \le k \le N-1 \qquad (4)$$

For efficient performance of the system, we add the logarithmic energy and compute the first difference (Δ). Taking logarithmic of energy compresses the dynamic range and makes the feature less variable to acoustic coupling changes. Moreover the human auditory system response is logarithmic in nature. Then cepstrum is calculated. It is the logarithmic form finally reverted into time. The output is referred to as MFCCs.

2. Peak detection:

Specifically, the aim of this work underlies in the development of an algorithm for peak detection possessing the following properties:

- Before the analysis, the user need not select any values concerning free parameters.
- Capability to examine the periodic as well as quasi-periodic signals for peaks.
- Attaining efficiency with regards to peak detection which is robust against noise (high/low frequency).

Thus, the peaks are found by use of an algorithm which confers the number of peaks along with the maxima and minima of the peak which act as one of the helpful features in mood detection.

IV. CLUSTERING

Clustering is basically a multiple objective problem requiring optimization. The most suitable algorithm for clustering as well as setting of parameters comprising values like a density threshold, distance function in use, the expected cluster number etc. is dependent upon the dataset as well as usage of the results intended.

Clustering on the basis of K-Means procedure originates a particular number of clusters that are not joint and are flat in nature i.e. non-hierarchical. It is quite suitable for formulation of clusters that are globular in nature. K-Means approach tends to have a numerical nature, is iterative, non-deterministic as well as unsupervised.

Algorithm Description- k means:

- K different points are placed in the space that is depicted by utilizing the objects which undergo the clustering process. Initial group centroids are demonstrated with the help of these points.
- The group with the centroid closest to the object is allocated to the later.
- Re-evaluation of the locations of K centroids takes place, after every object has been assigned to a group.
- The process mentioned in above the steps is repeated till the movement of the centroids ceased. This results in formation of groups comprising objects from which the calculation of the metric to be reduced is done.

This particular algorithm leads to the minimization of a target function which here is an error function that is squared. Eq. 5 showing the target function is given below:

$$J = \sum_{j-1}^{k} \sum_{i-1}^{x} \left\| x_i^{(j)} - c_j \right\|^2 \qquad (5)$$

Where $\left\| \mathrm{x_i}^{(\mathrm{j})} - \mathrm{c_j} \right\|^2$ is a selected measure of distance separating a point of data from the centre of cluster reflecting the separation among the 'n' number of points of data and their specific cluster centres**.**

V. DECISION RULE

K-means clustering is an unsupervised classification method. The MFCCs [5] of a happy song and a sad song is extracted and classified using K-means clustering [2]. Hence a binary split is done with one cluster showing happy mood musical clip and another cluster shows the sad mood musical clip.

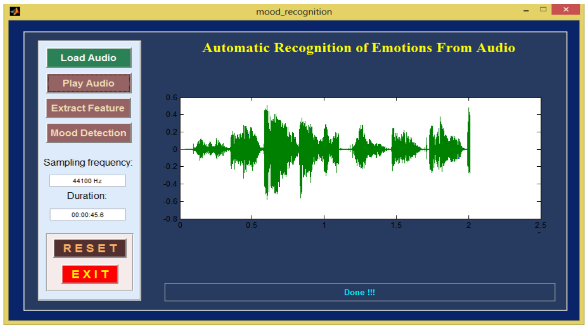
Another method of tagging a mood to the musical song is on the basis of peak difference and frame energy. A comprehensive study on the training set has been done and it is observed that sad mood music clips have peak difference less than 0.6000 and frame energy less than 57.0000. On the other side, the happy mood songs have peak difference greater than 0.8000 and energy greater than 57.0000. The confusion lies when the peak difference is between 0.6000 and 0.8000. To avoid this ambiguity the following algorithm is designed.

```
Step1: Calculate the peak difference and frame energy of an audio clip.
Step 2: if (peak difference < 0.6000)
            then tag the song with "sad" mood.
Step 3: if (0.6000< peak difference < 0.8000)
                    if (57.000< frame energy< 69.3000)
                    then tag the song with "Happy" mood.
        else
        tag the song with "sad" mood.
Step 4: if (0.8000 < peak difference < 1.0000)
                    if (frame energy < 55.000)
                    then tag the song with "sad" mood.
        else
        tag the song with "Happy mood".
```

The above mentioned algorithm acts as a decision rule for tagging a mood to the musical clip.

VI. SIMULATION AND RESULTS

Experiments have been performed on many songs of Indian music for training and testing. The songs with duration of 45s are chosen, each having a frame length of 60ms. For simulation a GUI was developed. The system was trained with 30 songs and testing is observed on over 100 songs. The simulation results for one audio clip are shown below. In pre-processing the raw audio clip is transformed into .WAV format with stereo type and 16 bit PCM. The sampling rate is chosen to be 44.1 KHz. After pre-processing the audio clip is played and displayed. Fig.2 describes the snapshot of the GUI designed describing the plot of an audio clip.


mood_recognition

Automatic Recognition of Emotions From Audio

Sampling frequency: 44100 Hz

Duration: 00:00:45.6

Done !!!

Fig. 2: Plot of a sample audio clip in the GUI developed

After processing the signal with Fourier Transform the absolute magnitude is calculated and can be shown by Fig. 3.

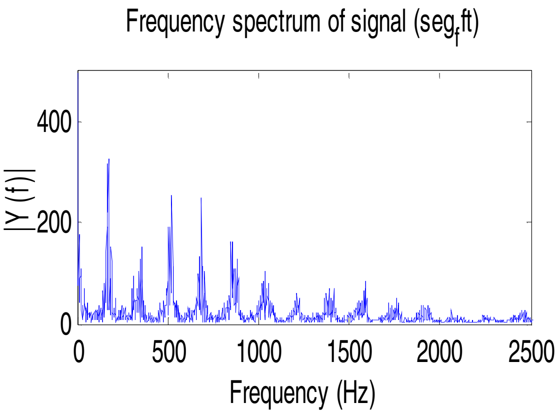


Fig. 3: Frequency spectrum of the audio clip

Now, the next step is to extract features from the processed audio clip. The MFCCs along with frame energy of the respective frames is extracted. A frame of an audio clip and its corresponding spectrogram defining its frame energy is shown in Fig 4. The third feature, Peak difference is calculated by passing the processed audio clip from an edge detection filter. The edge detection filter provides the maximum and the minimum peak values of the signal. The envelope of the edge detection filter is shown in Fig. 5.

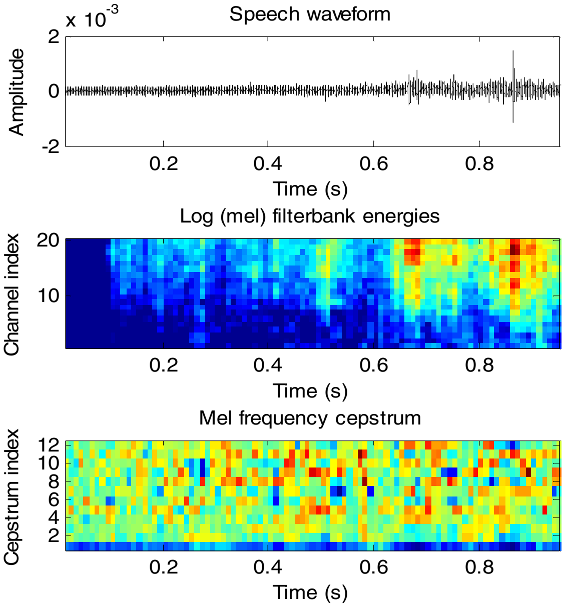


Fig. 4: Audio waveform and its Spectrogram

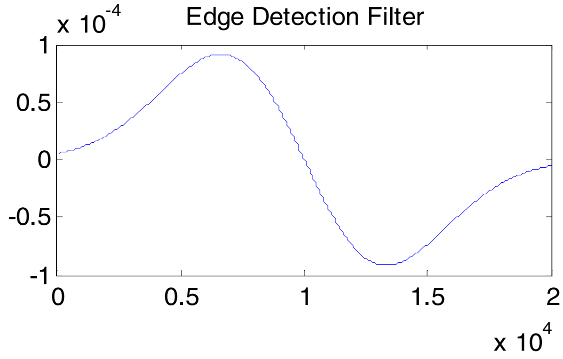


Fig. 5: Envelope of the signal from edge detection filter

Now the K-means clustering is applied which in turn produce three clusters of: blue squares (sad) and the red diagonals (happy). These clusters are shown in Fig 6.

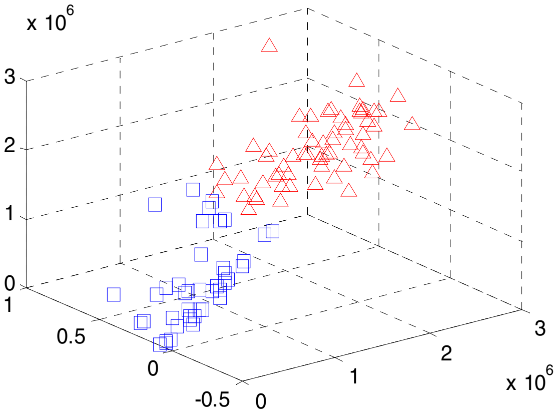


Fig 6: Clustering Analysis by K-means

The silhouette plots for the happy and sad clusters are shown below in Fig. 7.

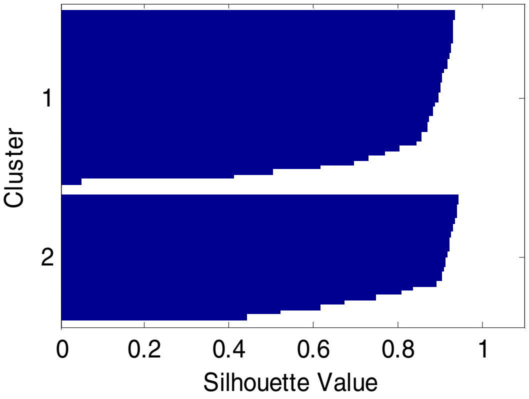
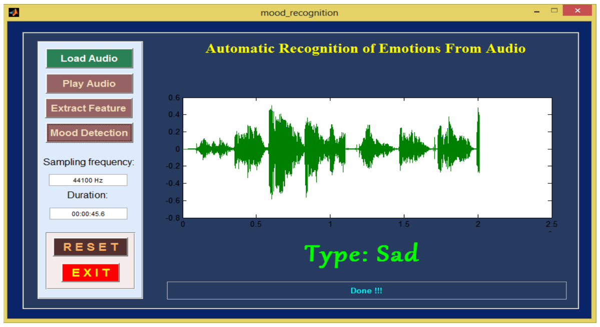


Fig. 7: Silhouette Plot

In testing phase, the thresholding value is selected on the basis of their peak difference value and frame energy. It is observed that the sad songs have peak difference of about 0.6000 and their frame energies less than 57.0000. On the other side, the happy songs have average peak difference a of 0.8200 and their frame energies are higher than 57.0000. After clustering and thresholding the mood of the song is detected and displayed. For the same audio clip the mood detected is "sad" and displayed in Fig. 8. Table 1 shows few audio clips from the training set and Table 2 shows the testing phase with 20 different audio clips.


mood_recognition

Automatic Recognition of Emotions From Audio

Sampling frequency: 44100 Hz

Duration: 00:00:45.6

Type: Sad

Done !!!

Fig. 8: "SAD" mood detected for an audio clip

Table 1: Few Audio Clips from the Training set with their peak difference and Frame energy.

S.No	Audio Filename	Duration of Audio clip	Mood Type	Peak Difference (Max. to Mini.)	Frame Energy (from MFCCs)
1	Audio Clip 1	45 seconds	Sad Mood	0.1161	16.8476
2	Audio Clip 2	45 seconds	Sad Mood	0.5038	69.0791
3	Audio Clip 3	45 seconds	Happy Mood	0.8025	69.4012
4	Audio Clip 4	45 seconds	Sad Mood	1.0000	49.6547
5	Audio Clip 5	45 seconds	Sad Mood	0.6825	52.8897
6	Audio Clip 6	45 seconds	Happy Mood	0.7038	59.5052
7	Audio Clip 7	45 seconds	Happy Mood	1.0000	60.7530

Table 2: Few audio clips showing classification of the songs on the basis of their mood with threshold value For the peak difference is 0.6000 and for frame energy the threshold is 57.0000

S.No	Audio Filename	Duration of Audio clip	Peak Difference	Frame Energy (from MFCCs)	Mood Detected	Mood Check
1	Audio Clip 1	45 seconds	1.0000	67.4322	Happy Mood	Correct
2	Audio Clip 2	45 seconds	0.8952	56.8521	Happy Mood	Correct
3	Audio Clip 3	45 seconds	0.9199	62.3548	Happy Mood	Correct
4	Audio Clip 4	45 seconds	0.6413	72.7458	Happy Mood	Incorrect
5	Audio Clip 5	45 seconds	0.8907	72.9514	Happy Mood	Correct
6	Audio Clip 6	45 seconds	0.8330	73.6589	Happy Mood	Correct
7	Audio Clip 7	45 seconds	0.9252	70.2564	Happy Mood	Correct
8	Audio Clip 8	45 seconds	1.0000	70.4368	Happy Mood	Correct
9	Audio Clip 9	45 seconds	0.6723	70.3658	Sad Mood	Incorrect
10	Audio Clip 10	45 seconds	0.6211	19.2458	Sad Mood	Correct
11	Audio Clip 11	45 seconds	0.6358	49.1149	Sad Mood	Correct
12	Audio Clip 12	45 seconds	0.2507	29.9581	Sad Mood	Correct
13	Audio Clip 13	45 seconds	0.1118	16.2894	Sad Mood	Correct
14	Audio Clip 14	45 seconds	0.4589	55.2378	Sad Mood	Correct

The accuracy is calculated as:

$$\text{Accuracy} = \left(\frac{\text{Total Number of Correct Samples}}{\text{Total Number of Testing Samples}}\right) \times 100$$

$$= \left(\frac{91}{100}\right) \times 100 = 90\%$$

VII. CONCLUSION

In this paper, the comprehensive set of audio features is used for the classification of the Indian music. The MFCCs, Frame energy and peak difference are the features of interest. The proposed scheme turned out to be much better as compared to not just other decision tree based algorithms but other classification algorithms as well. Classification is conducted by decision level fusion method. The result provided contentment as the accuracy was around 90% leading to good outcomes. The system may be widened for including different genres of songs such as Carnatic music, Hip-hop, Indian classical music , etc. by making suitable variations indulging acoustic features as well as techniques used for classification.

REFERENCES

[1] Tsunoo, E., Akase, T., Ono, N., Sagayama S., "Musical mood Classification by rhythm and bass-line unit pattern analysis", in Proc. IEEE International Conference on Acoustics, Speech and Signal Conference on Distributed Frameworks for Multi-media Applications, 2010, Yogyakarta, page(s) 1-5.

[2] Dewi K.C., Harjoko. A., "Kid's Song Classification Based on Mood Parameters Using K-Nearest Neighbor Classification Method and Self Organizing Map.", in Proc. International conference on Distributed frameworks for multimedia applications, 2010, page(s):1-5.

[3] Hunter, P. G.; Schellenburg, E. G., & Schimmack, U. "Feelings and perceptions of happiness and sadness induced by music: Similarities, differences, and mixed emotions". Psychology of Aesthetics, Creativity, and the Arts volume 4, 2010, page(s): 47–56.

[4] Ali, S. O.; Peynircioglu, Z. F. "Intensity of emotions conveyed and elicited by familiar and unfamiliar music". Music Perception: An Interdisciplinary Journal 27, 2010, page(s): 177–182.

[5] O. Lartillot and P. Toiviainen. Mir in matlab (ii): A toolbox for musical feature extraction from audio. In Proceedings of the International Conference on Music Information Retrieval, 2007, Vienna, Austria, page(s): 78-83.

[6] Dalibor Mitrovic, Matthias Zeppelzauer, Horst Eidenberger, "Analysis of the Data Quality of Audio Descriptions of Environmental Sounds", Journal of Digital Information Management, volume 4, 2007, Grienchland , page(s) 70-79

[7] McEnnis, D., McKay, C., Fujinaga, I., Depalle P., "jAudio: A feature extraction library", in the Proc. International Conference on Music Information Retrieval, 2005, page(s): 22-28.

[8] Li T., Ogihara, M., "Detecting emotion in music", in the Proc. International Symposium on Music Information Retrieval, Washington D.C., USA, 2003, page(s): 157-163.

[9] Han, B., Rho, S., Dannenberg, R. B., Hwang E., "Smers: Music emotion recognition using support vector regression", in Proc. 10th Intl. Society for Music Information Conf., Kobe, Japan,2010, page(s): 223-229.

[10] J. Skowronek, M. F. McKinney, and S. van de Par."A demonstrator for automatic music mood estimation" In Proceedings of the International Conference on Music Information Retrieval, Vienna, Austria, 2007,page(s): 169-175.

[11] Russell J. A., "A circumplex model of affect",Journal of Personality and Social Psychology, 1980, 39: page(s):1161-1178.