A New Recognition Method for Visualizing Music Emotion

Van Loi Nguyen[1], Donglim Kim[2], Van Phi Ho[3], Younghwan Lim[4]
[1,2,4]Department of Digital Media, Soongsil University, Korea
[3]Department of Computer Science and Engineering, Soongsil University, Korea

Article Info

Article history:

Received Feb 12 2017
Revised Apr 20, 2017
Accepted May 4, 2017

Keyword:

Feature extraction
Music emotion recognition algorithm
Music information retrieval
Music mood detection

ABSTRACT

This paper proposes an emotion detection method using a combination of dimensional approach and categorical approach. Thayer's model is divided into discrete emotion sections based on the level of arousal and valence. The main objective of the method is to increase the number of detected emotions which is used for emotion visualization. To evaluate the suggested method, we conducted various experiments with supervised learning and feature selection strategies. We collected 300 music clips with emotions annotated by music experts. Two feature sets are employed to create two training models for arousal and valence dimensions of Thayer's model. Finally, 36 music emotions are detected by proposed method. The results showed that the suggested algorithm achieved the highest accuracy when using RandomForest classifier with 70% and 57.3% for arousal and valence, respectively. These rates are better than previous studies.

Copyright © 2017 Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

Van Loi Nguyen,
Department of Digital Media,
Soongsil University,
369 Sangdo-ro, Dongjak-gu, Seoul, 06978, Korea.
Email: vanloiktkt@yahoo.com

1. INTRODUCTION

Music is a popular entertainment in our daily life, especially in the digital age. The emotional power of music is conceivably greater than the emotional power of any other medium. Music can bring us to tears and lull a baby to sleep. Its influences are felt clearly, but the ways it uses to touch our hearts are more indefinable [3]. Visualization techniques create a fitting environment for presenting emotion in music. Their capacity for creativeness, flexibility, and multidimensionality enables the interpretation of abstract data. Besides that, the primary motive to listen to music is to feel emotions. Therefore, the music performed with a visualizer helps audiences enhance the musical experience. Emotion visualization in music can be divided into two major stages: detecting emotion and visualizing emotion. Each detected mood is presented by a pattern of visualization media such as responsive behaviors of virtual character [13], photos [5], and a colored bar graph with text [10]. It is important to gain enough number of visual patterns to visualize music emotion more effectively. It means that a larger number of emotions must be appropriately recognized for visualization. In this paper, we propose an emotion recognition algorithm for music emotion visualization. As a result, 36 emotions are detected based on the level of arousal and valence in Thayer's model [2].

Recently music emotion recognition (MER) has become an active research topic that has been addressed by categorical approaches or dimensional approaches. Essential to the categorical approach is the concept of basic emotions or emotion clusters (e.g., happy, angry, sad, and fear) and applies machine learning techniques to create a training model. The idea is that there is a limited quantity of innate and common emotion categories from which all other emotion classes can be derived. The notion of basic emotions is diversified; different studies have employed different sets of basic emotions. Moreover, there is no distinction between songs grouped in the same category, even if there are obvious differences in terms of how strong the evoked emotions are. The major drawback of the categorical approach is that primary emotion classes are too small in comparison with the richness of music emotion perceived by humans. This approach is rarely used in emotion visualization because of the limited number of emotions (from 4 to 7 emotions). To solve this problem, our research purpose is to increase the number of detected music emotions for applications of music emotion visualization.

While the categorical approach aims mainly at the characteristics that distinguish emotions from one another, the dimensional approach to emotion conceptualization focuses on identifying emotions based on their positions on a small number of emotion "dimensions" with named axes, which are intended to correspond to internal human representations of emotion. These internal emotion dimensions are found by analyzing the correlation between affective terms. Until now, a few emotion models have been proposed in psychology and physiological sciences. One of the most known models was suggested by Thayer in 1989 as shown in Figure 1. Thayer's model is based on two basic and effective parameters, music energy and music pleasure, these parameters are also known as arousal and valence, respectively.

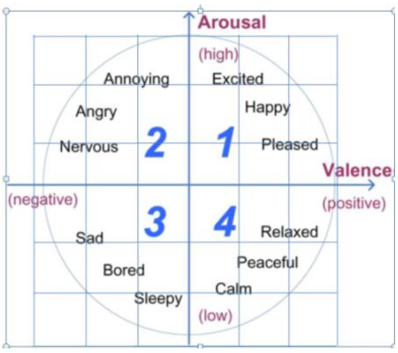

Figure 1. Thayer's Model for the Emotional Plane (also named Russel's model)

With the continuous approach, the arousal and valence values of each music sample are computed and the music sample is viewed as a point in the emotion plane. It is impossible to perform a music emotion by a visual pattern because we have to need innumerable visual ones. Therefore, the number of recognized emotions must be finite and big enough for emotion visualization. To that end, we divided Thayer's model into discrete emotion sections based on the level of arousal and valence. Emotion recognition is conducted on both dimensions arousal and valence. A detected music emotion belongs one of these emotion sections.

In summary, this paper offers the following original contributions which relate to the MER / MIR (music information retrieval) field:

a. The first study combining dimensional and categorical approaches in MER based on the level of valence and arousal;
b. The accuracies of proposed method are markedly improved comparing with previous work, especially for valence recognition of emotion.

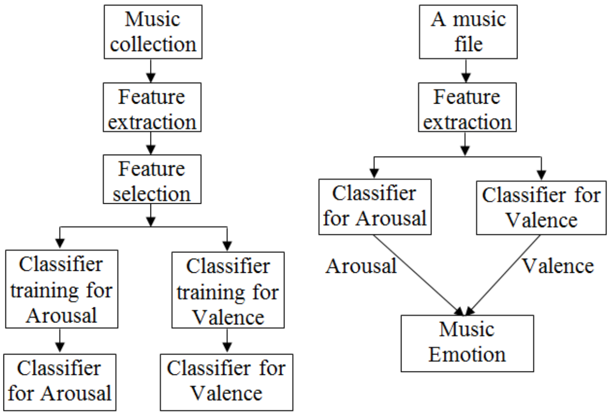

Figure 2. The Diagram of the Proposed Emotion Recognition Algorithm

Figure 2 presents the whole process of the music emotion recognition used in our work. The suggested system can be divided into two stages: building training model and classifying emotion. Firstly, 300 music clips are collected from All Music Guide (AMG) [19]. Next, the acoustic features are extracted from music dataset. After that, feature selection method is used for selecting only the most relevant features. The feature dataset is used to create 2 training models for arousal and valence. The learned models are then employed to predict the level arousal and valence of music files.

The rest of this paper is organized as follows. Related work is shown in section 2. Section 3 shows a music emotion recognition algorithm. Experimental results are presented in section 4. Finally, section 5 provides conclusions and future work.

2. RELATED WORK

One of the first studies in the categorical approach of MER was conducted by Feng, et al. in 2003 [6]. They classified music emotions into four categories consisting of happiness, sadness, anger, and fear. Tempo and articulation were extracted from 223 pop music segments. These features are employed as the input values of the neural networks. The average accuracy of the algorithm achieved at 67%. Late on, Yang, , et al. utilized a fuzzy k-NN classifier and a fuzzy nearest-mean classifier to detect four classes of music emotion from 243 pieces of modern pop music [15]. The best accuracy of the method reached 78.33%.

In dimensional approach, Korhonen, et al.applied the system identification technique to model the music emotion as a function of 18 musical features [8]. The accuracy of the method in terms of the R2 statistics achieved at 78.4% for arousal and 21.9% for valence. In another study, Yang, et al. formulated MER as a regression problem to predict the arousal and valence values of each music sample directly [14]. The best performance reached 0.58 for arousal and 0.28 for valence in terms of the R2 statistics by using support vector machines as regressors.

The viability of an MER system largely lies in the accuracy. From a review of the existing literature, it is clear that the accuracy of emotion recognition in previous work is not high especially for valence prediction, so there is plenty of room for improvement. In categorical approach, the accuracy depends on the number of detected emotions, music dataset, feature set, feature selection method, and classification algorithm. In this study, we used a bigger music dataset and feature set comparing with many previous studies. Finding optimal parameters of classification algorithms was executed through experimental studies. The method was evaluated with several machine learning techniques and its accuracy archived at 70% for arousal and 57.3% for valence. These results are a significant improvement comparing to existing work.

3. MUSIC EMOTION RECOGNITION ALGORITHM

3.1. Data Collection

There is not a standard database for recognizing music emotion; therefore most researchers have to compile their own. A large-scale database covering all kinds of music types and genres is desirable. There are many factors that impede the development of a common database. First, there is still no consensus on which emotion model or how many emotion categories should be. Second, due to the copyright issues, the audio files are needed for extracting music features relevant to emotion expression. Besides that, music clips are manually labeled, therefore the size of the music datasets in the previous work related to MER are usually not big enough [6], [15], [16].

300 music clips from AMG were utilized to create a training model in our experiment. AMG is a music company which uses moods for music organization and retrieval. This company uses up 288 emotion labels to describe the affective content of music, and these labels are assigned manually by musical experts [3]. This music collection is labeled with six levels of arousal and valence. As introduced in Table 1, the arousal level of Excited and Annoying is the highest while the highest valence level belongs to Pleased and Relaxed.

Table 1. The Level of Arousal and Valence in Thayer's Model

Arousal (Energy)		Valence (Pleasure)	
Level	Emotion	Level	Emotion
1	Sleepy, Calm	1	Nervous, Sad
2	Bored, Peaceful	2	Angry, Bored
3	Sad, Relaxed	3	Annoying, Sleepy
4	Nervous, Pleased	4	Excited, Calm
5	Angry, Happy	5	Happy, peaceful
6	Excited, Annoying	6	Pleased, Relaxed

A period 30 seconds of one track is used to normalize emotional variation because each track of music contains a range of different emotions. Furthermore, according to previous studies, a 30-second segment is a common choice, perhaps because it is the typical length of a chorus section of popular music and informative enough to retrieve the mood [3].

Music clip normalization is performed to deal with the amplitude differences of the music tracks in the database (i.e., some tracks have a higher volume than others). For the purpose, Cool edit pro is employed as an audio editing tool [20]. All 300 music clips are sampled at 44.1 kHz and 16 bits per sample.

3.2. Music Features

Previous research has indicated that music emotion is linked to music features. For example, sad music correlates with a slow tempo, whereas happy music is generally faster. The experience of music listening is multidimensional. Different emotion perceptions of music are usually associated with different patterns of acoustic cues. For example, while arousal is related to tempo (fast/slow), pitch (high/low), loudness (high/low), and timbre (bright/soft), valence is related to mode (major/minor) and harmony (consonant/dissonant). It is also noted that emotion perception is rarely dependent on a single music element but a combination of them. For example, loud chords and high-pitched chords may suggest more positive valence than soft chords and low-pitched chords, irrespective of mode. There are five feature groups presenting perceptual dimensions of music listening: energy, rhythm, temporal, spectrum, and harmony [3].

Energy Features: The energy of a song is often highly correlated with the perception of arousal. The perceived loudness can be measured by the dynamic loudness model of Chalupper and Fast implemented in PsySound [1]. We can also use it to extract 40 energy-related features including audio power (AP), total loudness (TL), and specific loudness sensation coefficients (SONE).

Rhythm Features

Rhythm is the pattern of pulses/notes of varying strength. It is often described in terms of tempo, meter, or phrasing. A song with a fast tempo is often perceived as having high arousal. Besides, flowing/fluent rhythm is usually related to positive valence, while firm rhythm is related to negative valence.

We can use Marsyas to compute the beat histogram of music and generate six features from it, including beat strength, amplitude and period of the first and second peaks of the beat histogram, and the ratio of the strength of the two peaks in terms of BPM (beats per minute).

Finally, the following five rhythm features proposed in [11] have also been shown relevant to both valence and arousal perception: rhythm strength, rhythm regularity, rhythm clarity, average onset frequency, and average tempo. All of them can be extracted using the utility functions in the MIR toolbox [9].

Temporal Features: We use SDT to extract temporal centroid, zero-crossing rate, and log attack time to capture the temporal quantity of music [4]. Zero-crossing rate, a measure of the signal noisiness, is computed by taking the mean and standard deviation of the number of signal values that cross the zero axis in each time window (i.e., sign changes). Temporal centroid and log attack time, on the other hand, are two MPEG-7 harmonic instrument timbre descriptors that describe the energy envelope.

Spectrum Features: Spectrum features are features computed from the STFT of an audio signal. One can use Marsyas to extract the timbral texture features including spectral centroid, spectral rolloff, spectral flux, spectral flatness measures (SFM), and spectral crest factors (SCF). These features are extracted for each frame and then by taking the mean and standard deviation for each second. The sequence of feature vectors is then collapsed into a single vector representing the entire signal by taking again the mean and standard deviation. Note that many of these features can also be extracted by the SDT. Mel-frequency cepstral coefficients (MFCCs) can be also extracted by SDT, the coefficients of the discrete cosine transform (DCT) of each short-term log power spectrum expressed on a nonlinear perceptual-related Mel-frequency scale, to represent the formant peaks of the spectrum.

Harmony Features: Harmony features are computed from the sinusoidal harmonic modeling of the signal. A larger number of natural sounds, especially musical ones, are harmonic each sound includes a series of frequencies at a multiple ratio of the lowest frequency, named the fundamental frequency f0. We can use the MIR toolbox to generate two pitch features (chromagram center and salient pitch) and three tonality features (mode, key clarity, harmonic change).

3.3. Feature Extraction and Feature Selection

3.3.1. Feature Extraction

In this work, Sound Description Toolbox, MIR Toolbox, Marsyas, and PsySound were used to extract features from the audio clips.

The SDT extracts a quantity of MPEG-7 standard descriptors and other features from WAV audio files.

The MIR Toolbox framework is an integrated set of functions written in MATLAB, that are specific to the extraction and retrieval of musical information such as pitch, timbre, tonality and others. This framework is widely used and well documented, providing extractors for a high quantity of both low and high-level audio features.

PsySound3 is a toolbox written in MATLAB for the analysis of sound recordings using physical and psychoacoustical algorithms. It does precise analysis using standard acoustical measurements, as well as implementations of psychoacoustical and musical models such as sharpness, loudness, roughness, fluctuation strength, rhythm, pitch, and running interaural cross-correlation coefficient (IACC).

Marsyas (Music Analysis, Retrieval, and Synthesis for Audio Signals) is a framework created for audio processing with concrete emphasis on MIR applications. Written in highly optimized C++ code, it stands out from the others due to its performance, one of the main reasons for its adoption in a variety of projects in both academia and industry. Some of its pitfalls are the complexity and the lack of some features considered relevant to MER [7].

A total of 397 acoustic features were extracted, 187 using SDT, 174 using Marsyas, 30 using MIR Toolbox and 6 with PsySound 3. Regarding the analysis window size used for frame-level features and hop size, all default options were used (512 samples for Marsyas and 0.05 seconds for MIR Toolbox). These features are then transformed in song-level features by calculating mean, variance, kurtosis and skewness. A small summary of the extracted features and their respective framework is given in Table 2.

Table 2. List of Audio Frameworks used Foreature Extraction

Framework	Features	Total
SDT	Audio Fundamental Frequence, audio power, Audio spectrum Spread, Autocorrelation, log attack time, Mel-frequency cepstral coefficients (MFCCs), specific loudness sensation coefficients (SONE), Spectral centroid, spectral flatness measures (SFM), spectral rolloff, temporal centroid, total loudness, Zero-crossings	187
Marsyas	Beat histogram, pitch histogram, spectral crest factors (SCF), spectral flux	174
MIR Toolbox	Average onset frequency, average tempo, chromagram centroid, harmonic change, inharmonicity, irregularity, key clarity, musical mode, rhythm clarity, rhythm regularity, rhythm strength, roughness, salient pitch	30
PsySound 3	Dynamic loudness, sawtooth waveform inspired pitch estimate, sharpness	6

The feature vectors have to be standardized after the feature extraction because some features (with bigger values) have more influence when creating the training model. This step is important, especially for classification algorithms that do not have a mechanism for feature standardization. As a result, the distribution of the values of each feature is with standard deviation equal to 1 and mean equal to 0. And the values for each feature are on the same scale from -1 to 1.

3.3.2. Feature Selection

From the viewpoint of machine learning, feature selection is the process of choosing a subset of relevant features because irrelevant or redundant ones can lead to an inaccurate conclusion. To address this problem, a number of music features need to be extracted and then use a feature selection algorithm to find out good features. The feature selection is employed to find the optimal feature subset that gives the maximal recognition accuracy and keeps the feature dimension minimal. In addition, it lowers down the computational time of the experiment.

For simplicity and effectiveness, Attribute Evaluator = “CfsSubsetEval” and Search Method=”BestFirst” are used in our work. As a result, 28 attributes for arousal and 27 attributes for valence are selected from 218 attributes and 323 attributes, respectively.

3.4. Training Process

In our experiments, WEKA is employed for building training models. It is a machine learning tool that uses approach for parameter optimization of classification algorithms. To increase the accuracy of the chosen algorithm, WEKA finds the near optimal parameter setting in the huge space of algorithm parameters by using intelligent optimization functions [17].

After the features are extracted, standardized and selected, they are utilized to form the feature vector database. Each music clip in the database is an instance, i.e., feature vector, used for classification. Supervised classification algorithms are used in our experiments because each instance is labeled with the suitable arousal level or valence level. We conducted thorough experiments with each of the classification algorithms, and once we selected the one with the highest recognition accuracy, we further enhanced its accuracy with WEKA. These classification algorithms are employed to train two distinct models, the one for valence and the other for arousal. For testing options, we used 10-fold cross-validation technique which used as a WEKA pre-established parameter for performance evaluation of our chosen classifier algorithm. The training model and estimation are presented in Figure 3.

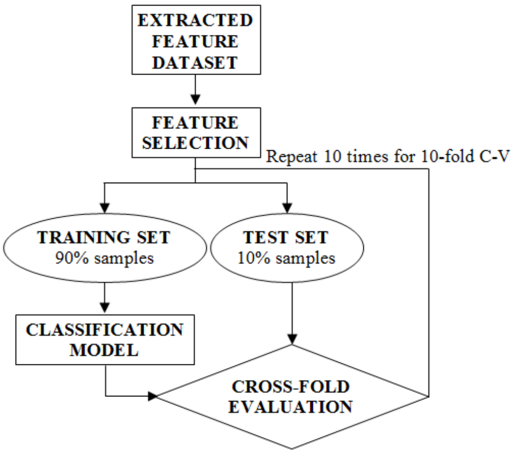

Figure 3. Flow-chart of tTaining Model and Evaluation

In machine learning, three are common evaluation metrics that were analyzed: the recall, precision, and accuracy. The below formulas define each of the metrics, where Q can be any level of arousal or valence that we are trying to detect (level 1, level 2, etc.).

$$\text{recall} = \frac{\text{No. of correctly recognized levels labeled as Q}}{\text{No. of all the levels labeled as Q}} \qquad (1)$$

$$\text{precision} = \frac{\text{No. of correctly recognized levels labeled as Q}}{\text{No. of all the levels recognized as Q}} \qquad (2)$$

$$\text{accuracy} = \frac{\text{No. of correctly recognized levels of all types}}{\text{No. of all the levels}} \qquad (3)$$

4. EXPERIMENTAL RESULTS

We carried out various experiments in the following order. Firstly, 397 acoustic features were extracted from 300 labeled music clips by SDT, MIR Toolbox, Marsyas, and PsySound. As presented in table 4, these features were then divided into two feature subsets to build two training models for arousal and valence because many features are only highly correlated with the perception of valence or arousal.

Table 4. Two Feature Subsets using in the Experiments

	Feature group					Total
	Energy	Rhythm	Temporal	Spectrum	Harmony	
Features for arousal	Audio fundamental frequence, audio power, dynamic loudness, sharpness, SONE, total loudness.	Beat histogram, average onset frequency, average tempo, rhythm clarity, rhythm regularity, rhythm strength	Autocorrelation, log attack time, temporal centroid, zero-crossings	Spectral centroid, spectral rolloff, audio spectrum spread, irregularity, MFCCs, roughness, spectral flux	Chromagram centroid, salient pitch	218
Features for valence		Average onset frequency, average tempo, rhythm clarity, rhythm regularity, rhythm strength	Autocorrelation, log attack time, temporal centroid, zero-crossings	Audio spectrum spread, irregularity, MFCCs, roughness, spectral flux, inharmonicity, SCF, SFM	Harmonic change, key clarity, musical mode, pitch histogram, sawtooth waveform inspired pitch estimate, chromagram centroid, salient pitch	323

After that, we tested the accuracy of all classification algorithms which were installed in WEKA. Training model evaluations were conducted in a 10-fold Cross Validation. The accuracies of classifiers are significantly improved when using feature selection as shown in Table 5.

Table 5. The Results of Training Model Evaluation

	All features		
Classifier	Arousal	Classifier	Valence
Functions.simpleLogistic	66%	Trees.LMT	51%
Functions.SMO	66%	Functions.SMO	51.3%
Trees.RandomForest	66%	Trees.RandomForest	52.7 %
	Feature selection		
Classifier	Arousal	Classifier	Valence
Trees.LMT	67.7%	Meta.Bagging	55.3%
Functions.simpleLogistic	68%	Functions.SMO	55.7 %
Trees.RandomForest	**70%**	Trees.RandomForest	**57.3%**

Table 5 shows the evaluation results of 3 classifiers which achieved the top 3 accuracies. When using attribute selection techniques, many redundant or irrelevant features were removed. As a result, time for running experiment decreased significantly and the accuracies of classifiers increased from 2% to 5%. The results showed that Trees.RandomForest classifier achieved the best accuracy with 70% for arousal and 57.3% for valence. RandomForest is a meta estimator that suits a number of decision tree classifiers on various sub-samples of the dataset and employ averaging to better the predictive accuracy and control over-fitting.

The confusion matrix of classification using RandomForest is presented in Table 6 and Table 7. Moreover, we show the recall and the precision for each class (level of arousal or valence), and the overall accuracy. In Table 6, the highest precision and recall are attained for the class “level 5” and the lowest are achieved for the class “level 4”. It is clear that the classes “level 4” and “level 6” are often mixed by the classifier.

Table 6. Confusion Matrix for Arousal

10 fold cross-validation		PREDICTED CLASS FOR AROUSAL						Recall (%)
		Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	
	Level 1	44	0	4	1	0	1	88
	Level 2	0	47	1	0	2	0	94
REAL CLASSS	Level 3	15	0	24	8	0	3	48
	Level 4	6	0	16	15	0	13	30
	Level 5	0	3	0	0	47	0	94
	Level 6	0	0	4	13	0	33	66
Preciscion (%)		67.7	94	49	40.5	95.9	66	Acc=70%

In Table 7, the highest precision and recall are attained for the class “level 2” and the lowest are achieved for the class “level 1”. The class “level 6” is mixed with all other 6 classes.

Table 7. Confusion Matrix for Valence

10 fold cross-validation		PREDICTED CLASS FOR VALENCE						Recall (%)
		Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	
	Level 1	20	0	11	6	9	4	40
	Level 2	1	36	0	2	7	4	72
REAL CLASSS	Level 3	12	0	23	11	3	1	46
	Level 4	4	0	11	31	4	0	62
	Level 5	9	0	4	5	29	3	58
	Level 6	8	3	2	1	3	33	66
Preciscion (%)		37.0	92.3	45.1	55.4	52.7	73.3	Acc=57.3

The average accuracy of emotion recognition for arousal and valence is 63.7% with 6 emotion classes. According to the synthesis of the Music Information Retrieval Evaluation eXchange (MIREX), the highest reached classification accuracy in the Mood Classification Task (for five emotion classes) was

67.8% [12], [18]. The results of proposed method are better than previous studies [6], [15]. However, the accuracies between emotion classes have a big difference in our experiment.

5. CONCLUSION AND FUTURE WORK

In this paper, we used the machine learning techniques and Thayer's model to classify music emotions. Each music emotion can be recognized based on the level of arousal and valence. As a result, 36 detected emotions are enough for applications of music emotion visualization.

To improve the efficiency of emotion recognition in music, we used a big music dataset with 300 music clips. A large number of music features are extracted by extractors. Method selection is used to decrease the computing time and increase the accuracies of the proposed approach. With 6 emotion classes, the classification accuracies were 70% for arousal and 57.3% for valence with classifier RandomForest. As expected, these results are a clear improvement in comparison to previous studies.

In this study, there is a significant improvement for the accuracy of MER. However, the achieved results are still not very high because of the following reasons. Firstly, emotion perception is intrinsically subjective, and people can perceive different emotions for the same song. This subjectivity issue makes the performance evaluation of an MER system fundamentally difficult because a common agreement on the classification result is hard to obtain. Secondly, it is not easy to describe emotions in a universal way because the adjectives used to depict emotions may be ambiguous, and the use of adjectives for the same emotion can vary from person to person. Thirdly, it is still unexplained how music evokes emotion. What intrinsic element of music, if any, creates a concrete emotional response in the listener is still far from well-understood.

a. There are still some shortcomings in current research, such as low recognition rates, music emotion recognition is performed offline. In the following, some solutions for future works are recommended.
b. Based on proposed dimensional models in MER, the level of arousal and valence are flexibly divided according to concrete applications of emotion visualization.
c. Music emotion will be recognized in real time.
d. The system will use bigger music dataset and combine low-level audio features and mid-level features such as lyrics, chord progression, and genre metadata in order to achieve higher accuracy of MER.

REFERENCES

[1] Cabrera D, *et al.*, “Psysound3: A Program for the Analysis of Sound Recordings”, *Journal of the Acoustical Society of America*,123(5):3247, 2008 May.

[2] Thayer, *et al.*, “The Biopsychology of Mood and Arousal”, Oxford University Press, 1989.

[3] Yang YH, *et al.*, “Music Emotion Recognition”, CRC Press, 2011.

[4] Benetos E, *et al., “Large Scale Musical Instrument Identification”*, Proceedings of the 4th Sound and Music Computing Conference, Jul 2007.

[5] Chen Chin-Han, *et al.*, "Emotion-Based Music Visualization using Photos", *Advances in Multimedia Modeling, Springer Berlin Heidelberg*, pp. 358-368, Jan 2008.

[6] Feng Y, *et al.*, “Popular Music Retrieval by Detecting Mood”, Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 375-376, Jul 2003.

[7] J. Leben, *et al., “Declarative Composition and Reactive Control in Marsyas”*, Joint 40th International Computer Music Conference (ICMC) and 11th Sound and Music Computing conference (SMC), 2014.

[8] Kor honen MD, *et al.*, “Modeling Emotional Content of Music using System Identification”, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, pp. 588-99, Jun 2005.

[9] Lartillot O, , *et al., “A Matlab Toolbox for Musical Feature Extraction from Audio”*, In International Conference on Digital Audio Effects, pp. 237-244, Sep 2007.

[10] Laurier Cyril , *et al., “Mood Cloud: A Real-Time Music Mood Visualization Tool”*, Proceedings of the Computer Music Modeling and Retrieval, (2008).

[11] Lu L, *et al.*, “Automatic Mood Detection and Tracking of Music Audio Signals”, *IEEE Transactions on audio, speech, and language processing*, pp. 5-18, Jan 2006.

[12] Panda R, *et al., “Dimensional Music Emotion Recognition: Combining Standard and Melodic Audio Features”,* In Proceedings of the 10th International Symposium on Computer Music Multidisciplinary Research (CMMR), pp. 583-593, 2013.

[13] Taylor R, *et al., “Visualizing Emotion in Musical Performance using a Virtual Character”,* In International Symposium on Smart Graphics, Springer Berlin Heidelberg, pp. 13-24, Aug 2005.

[14] Yang YH, , *et al.*, “A Regression Approach to Music Emotion Recognition”, *IEEE Transactions on audio, speech, and language processing*, , pp. 448-57, Feb 2008.

[15] Yang YH, , *et al.*, *"Music Emotion Classification: A Fuzzy Approach",* Proceedings of the 14th ACM international conference on Multimedia, pp. 81-84, Oct 2006.

[16] Yeh CH, *et al.*, *"An Efficient Emotion Detection Scheme for Popular Music"*, In 2009 IEEE International Symposium on Circuits and Systems, pp. 1799-1802, May 2009.

[17] Weka. [Online] http://www.cs.waikato.ac.nz/ml/weka/

[18] http://www.music-ir.org/mirex/wiki/MIREX_HOME

[19] All music guide. [Online] http://www.allmusic.com/

[20] Cool edit pro. [Online] http://www.adobe.com/products/audition/

BIOGRAPHIES OF AUTHORS

Van Loi Nguyen, he received his Master of Engineering in Computer Science from the University of Danang, Vietnam in 2010. He is currently a Ph.D. candidate at Soongsil University, Korea. His research interests include multimedia, information retrieval, database, and software testing

Donglim Kim, she received a Master degree and a Ph.D. degree in Media Engineering from Soongsil University in 2011 and 2016 respectively. Her research interests include multimedia, motion engineering, content engineering.

Van Phi Ho, he received his BS and MS degrees in the Computer Science Department at Da Nang University in October 2004 and October 2009, respectively. He is currently a Ph.D. student in the School of Computer Science and Engineering, Soongsil University. His research interests include flash memory-based DBMSs and database systems.

Younghwan Lim, he received his Master from the Dept. of Computer Science of Korea Advanced Institute of Science & Technology in 1979, a Ph.D. degree from Northwestern University in 1985. From 1996 to the present, he has been a professor in Global School of Media, Soongsil University, Korea. His research interests include mobile solutions, multimedia, and creative engineering design.