International Journal of Fuzzy Logic and Intelligent Systems, vol. 11, no. 2, June 2011, pp. 124-128
DOI : 10.5391/IJFIS.2011.11.2.124

SYMMER: A Systematic Approach to Multiple Musical Emotion Recognition

Jaesung Lee, Jin-Hyuk Jo, Jae-Joon Lee and Daewon Kim*

School of Computer Science and Engineering, Chung-Ang University

Abstract

Music emotion recognition is currently one of the most attractive research areas in music information retrieval. In order to use emotion as clues when searching for a particular music, several music based emotion recognizing systems are fundamentally utilized. In order to maximize user satisfaction, the recognition accuracy is very important. In this paper, we develop a new music emotion recognition system, which employs a multilabel feature selector and multilabel classifier. The performance of the proposed system is demonstrated using novel musical emotion data.

Key Words : Music Emotion Recognition, Multilabel Classification, Feature Selection

1. Introduction

The recognition of musical emotions has emerged as a popular research area in music information retrieval and music psychology. It requires an effective technique to analyze and classify the emotional content of music. Traditional approaches to musical emotion recognition (MER) typically describe the music with a single emotion; they attempt to model MER problem as a single-label classification [1-6].

However, few researches have studied that music may express multiple emotions according to the section of given music and even by subjective feelings of listeners. It might start with excited and end with happy. To reflect this nature of musical emotions, a new method is required to handle the multiple emotion recognition. To achieve this, we addressed two issues in developing multiple MER systems (modeling and training/classifying emotional features measured from the music), and proposed a systematic approach to these issues.

The first contribution of this paper is to create and release a new, publicly available music data set that consists of musical features and multiple emotions; a music can be assigned maximum four different musical emotions. It would be helpful for researchers who have trouble developing MER system due to the rarity of multi-emotion data. For modeling, we employed Russell's emotion theory [7] and the MIR toolbox [8].

The second contribution is to develop new MER system that interprets the relation between musical features and multiple emotions. From the hundreds of musical features measured from the same music spectrum, but different in time intervals, we tried to identify significant features that are highly relevant to emotions, along with minimizing redundancy to others.

The organization of this paper is as follows. Section 2 introduces some milestone works of music emotion recognition and the musical feature selections. Section 3 presents a systematic musical feature selection approach to multiple musical emotion recognition. Experimental results from the application of two different feature selection methods to recognize multiple music emotion are presented in Section 4. Section 5 presents conclusions are.

2. Related Work

The goal of MER is to recognize the internal emotion of given music. To achieve the main goal of MER, first we need a dataset. The procedure of making a MER dataset is given below, it is mainly focused on representing physical properties of music itself [2]. Typically each music was assigned to each pattern, and music can be given differently according to each research, for example, entire music (total duration in playing time) or a segment of music (as known as a music excerpt).

Next the features were extracted from music signal. Due to the difficulty of handling temporal signal information, most of feature extraction techniques were based on some signal transformation techniques such as short-term fourier transform (STFT), mel-scale frequency cepstral coefficient (MFCC) and some novel heuristic musical property detectors. The signal transformation techniques were focused on spectral shape of given music signal that described by STFT, MFCC, ZCR (Zero-Crossing Rate) and so on. To apply signal transformation techniques, they divide these music signal into user-defined time intervals. To prevent information loss, these unit intervals were overlapped to near time intervals both previous interval and next interval. Then musical properties were conversed from signal to constant value on each unit interval.

Manuscript received Mar. 28, 2011; revised May. 18, 2011.
* Corresponding author
This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2010-0012885)

Final step of making data set is to collect the emotional response, also called class label in machine learning domain. Due to the subjective nature of music emotion, some researches annotated music emotion as multiple emotion [2, 9-10].

At the end, appropriate classifier (also called recognizer) can be used to recognize the musical emotion from musical features, for example, naive bayes and support vector machines. Some milestone works in MER domain introduced in next paragraphs.

Feng *et al.* [9], proposed a neural network classifier to detect music emotion. There intention of is to express the human perception of tempo and articulation in a implied manner. The structure and weight matrix of the neural network provide us with the knowledge about how people regard tempo as being "fast" or "slow" and articulation as being "staccato" or "legato" in fuzzy quantity with output value ranged [0 1]. They employed the emotion "score" instead of explicit judgment such as "fear" or "happy" to depict music emotion. If score is larger than 0.5 for a specific output node, the piece of music is in the corresponding emotion, if all score is less than 0.5, the representing emotion by the ouput node with largest score value is assigned to the music piece.

Trohidis *et al.* [10], examined multi-label classification of music emotion. They employed the Support Vector Machine (SVM) to several experimental designs such as Label Powerset (LP), Binary Relevance (BR) and RAKEL[11]. They examined that the effect of multiple emotion representation. At the last, they showed the predictive performance of the four competing multi-label classification algorithm using a variety of measures. Because the correlation among emotions, they noticed that RAKEL dominates the other algorithms in almost all measures.

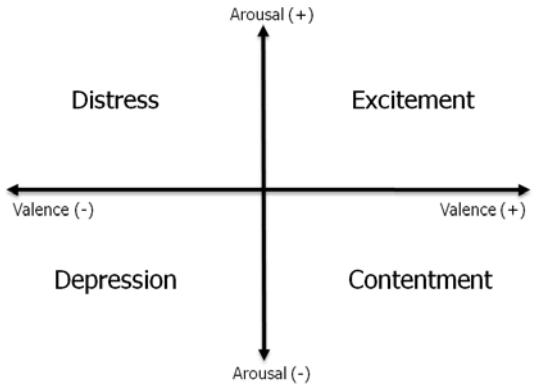


Fig. 1. The Russell's emotion model; emotion is represented as a point defined by arousal and valence coordinates.

3. Proposed Method

3.1 Modeling of Multiple Emotions

A successful modeling of multiple emotions from music data plays key roles in MER problem; however, it is hard to obtain a data set composed of musical features and emotions, so we assembled the data set ourselves. To this end, we collected 100 songs from five genres (Ballads, Hip-Hop, Dance, Trot, and Rock) and had them subjectively labeled by 500 participants through an on-line annotation system. For each song, we applied the MIR toolbox that offers an integrated set of functions to extract musical audio features. The extracted features fall into six types: dynamics, fluctuation, rhythm, spectral, timbre, and tonal features.

In this modeling of the emotional content of music, we employed the most widely used Russell's emotion theory. In Russell's model, a state of human emotion is represented as a point in a two-dimensional emotion space; in Fig. 1, it is defined by two fundamental emotional basis, arousal and valence coordinates. Hence, the multiple emotions can be plotted as multiple points in the emotion model. For simple description of multiple emotions for each music, in this paper we used the four sub-planes (zones) in Russell's model as emotion labels. The first plane, positive arousal and positive valence labeled by $l1$={+,+}, represent Excitement feeling. Other three planes can be labeled as Distress $l2$={+,-}, Depression $l3$={-,-}, and Contentment $l4$={-,+}. We assumed that the perceived emotion may distribute equally all of planes, if a music does not evoke any emotion. Since each song is labeled by many participants who may perceive different emotions from it, we represent the multiple emotions of a song as a set of aggregated labels over the four emotion planes if an emotion is labeled by more than 25% of participants. For example, a music that gave Excitement(52%), Distress(4%), Depression(12%) and Contentment(32%) feeling to participants, then this song assigns to $l1$ and $l4$ simultaneously.

As a result, we were able to create and release a new multi-emotion data set with 864 musical features for 100 different songs categorized into one or more of four emotions. It is publicly available from Softwares section of http://ai.cau.ac.kr/.

3.2 Selection of Significant Musical Features

As Yang et al. [1] pointed out, hundreds of musical features are not typically of equal importance or quality so that irrelevant or redundant features may degrade recognition accuracy and learning time. Most of traditional researches selected good features based on expert's suggestion. One straightforward way to select the significant features is to identify the emotion relevant features and their relation to other features for multiple emotions. For its simplicity and effectiveness, we adopted the work of Peng et al. [12] to select musical features with both the highest relevance to the emotions and the lowest redundancy between them.

Let us define a data set that contains feature set F, emotion set L and is composed of n songs. The feature set F contains k musical features, F={ $f1$, … , fk }. The emotion set L contains q emotions, L={ $l1$, … , lq }. Suppose we now have St, the feature subset calculated from t-th step.

The presented method selects the next feature *fi* from *F-St* and add it *St+1*; *fi* is mutually far away from preselected feature subset *St* while still having a high correlation to emotions. To implement this, we used two criteria defined as relevance *D* and redundancy *R*. A feature *fi* is selected when *fi* is maximizing correlation between *St+1* and *L*, called relevance; *D*.

$$D(S,L) = \frac{1}{|S|}\sum_{f_i \in S} M(f_i;L) \tag{1}$$

where

$$M(f_i;L) = \max_{l_j \in L} I(f_i;l_j) \tag{2}$$

is a correlation score between *fi* and emotion set *L*,

$$I(f_i;l_j) = \sum\sum P(f_i,l_j)\log\frac{P(f_i,l_j)}{P(f_i)P(l_j)} \tag{3}$$

is a mutual information value between *fi* and *lj*.

However, *fi* may be mutually correlated with the preselected features in *St*, therefore, *St* should minimize the following redundancy criterion *R*:

$$R(S) = \frac{1}{|S|^2}\sum_{f_i,f_j \in S} I(f_i;f_j) \tag{4}$$

where *I*(*fi*;*fj*) is a mutual information value between *fi* and *fj*. To combine *D* and *R*, we exploited the operator *Φ* defined in [12], optimizing *D* and *R* simultaneously.

$$\max \Phi(D,R) \tag{5}$$

where *Φ=D-R*. An incremental search algorithm selects feature *fi* that is maximizing *Φ*. It is interesting to note that the minimizing redundancy scheme make force to select features balancely in the view point of relevance of each emotion.

4. Experimental Results

4.1 Performance of Emotion Recognition

To simulate the recognition accuracy of the proposed method, we employed the parallel naive bayes (PNB) classifier and well-known evaluation measures, which are popularized in multi-label classification problems [11]. PNB was trained same music data with each emotion, thus *q* independent naive bayes classifier was used to recognize the existence of each emotion in a song. And then one song was held out as an independent test set, a technique which is called leave one out cross validation. The rest of music data set was used to train PNB. These training and test process was repeated *n* times, where *n* is the number of music in given data set. We conducted PNB-based MER experiments using two types of feature selections for musical emotions: (1) the well-known $X2$ statistic (CHI) as the baseline method (CHI+PNB) (2) the proposed method (Proposed+ PNB); they were evaluated using four measures: 1-hamming loss, precision, recall, and macro F1.

Table 1 shows the recognition performance using feature subsets that are selected by each method. The proposed method showed superior recognition performance than CHI+PNB in the top 80 features (10% of the original feature set) over all the evaluation measures. For example, the 1-hamming loss score was 0.84 by PNB+Proposed, while that was scored 0.72 when CHI+PNB was applied. Thus we can see the performance is dominating 12% when it compared to CHI.

Table 1. Comparison results of Parallel Naive Bayes (PNB) performance when applied to 80 features (10% of the original feature set) selected by CHI and the proposed method respectively.

	Evaluation Measure			
	1-H.Loss	Precision	Recall	Macro F1
CHI+PNB	0.7150	0.5533	0.5417	0.4987
Proposed+PNB	**0.8425**	**0.7900**	**0.8733**	**0.7556**

Fig. 2 shows the overall recognition performance of the two methods in terms of 1-hamming loss. We can see that the proposed method leads to better results over the entire range of the size of feature subset. The growth of the 1-hamming loss score of the Proposed+PNB is faster than that of the CHI+PNB especially when the size of feature subset is small. It is evident that the proposed method achieves better recognition accuracy by selecting more compact and relevant musical features. The best 1-hamming loss score of the Proposed+PNB is 0.85 when 150 features are selected.

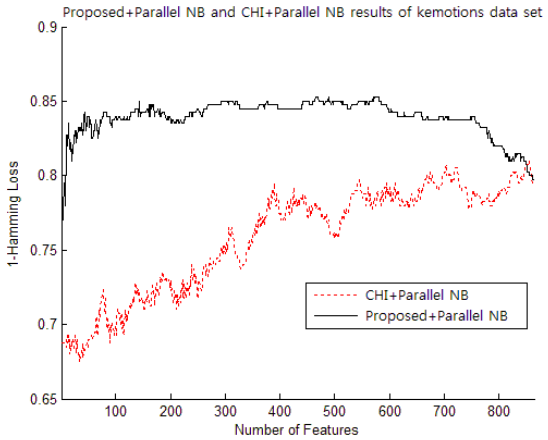


Fig. 2. Overall recognition performance of 1-hamming loss score when Proposed+PNB and CHI+PNB are performed to kemotion data set.

In addition, we tested the potential of the proposed method when applied to other domain data set that is composed of low-level information features and high level multiple semantics labels. Fig. 3 shows the overall recognition performance of the Scene data set [13] where some labels such as beaches, sunsets, or parties. As shown in Fig. 3, the proposed method selected effective features; the best

1-hamming loss score is measured when 15 features are selected. The best 1-hamming loss score of the Proposed+PNB is 0.83.

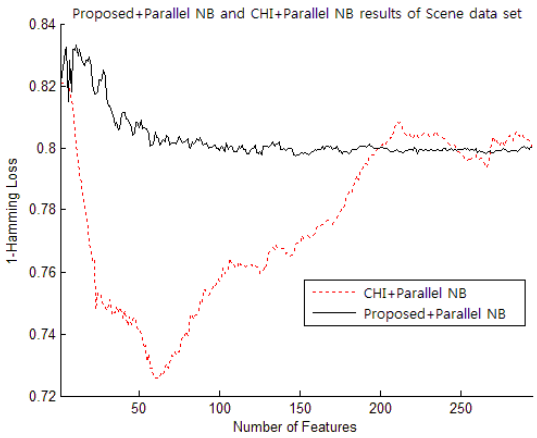


Fig. 3. Overall recognition performance of 1-hamming loss score when Proposed+PNB and CHI+PNB are performed to Scene data set.

A reason why Proposed+PNB dominates CHI+PNB is that the CHI+PNB does not consider the redundancy among selected features. CHI+PNB selected features individually according to the information of each feature and multiple class label, and then selected features according to the mount of information. However, as Peng *et al*. pointed out that "m best features are not the best m-features.", redundant features can be selected by CHI+PNB. This feature selection process may defect the compactness of selected feature subset and confuse recognizers. The proposed method consider the redundancy between a selected features and newly selected feature by Φ operator, and redundant feature does not selected even if this feature have much of information about multiple class label.

4.2 The Best Emotion-Relevant Features

We analyzed the relations between the selected musical features and emotions by the association rule learning. Table 2 shows the emotion, Arousal-Valance Status, and the name of the selected feature. For simplicity, we extracted two relations based on the the best support and confidence value for each emotion from the 80 features selected by the proposed method. For example, the first relation in Table 2 means that the presence of feature K1SS is significant to express the emotion Excitement; if input music gets positive in the feature K1SS, then it is recognized into Excitement, and if input music gets positive in the feature TM2S then this music may give Contentment.

Moreover, we investigated the co-occurrence ratio between the top ranked features and their related emotions (Fig. 4); we cut the bar-graph off if its ratio was lower than 0.5. We see that Excitement is positively correlated to K1SS and negatively correlated to either TM2F and TM2S. If K1SS and D1DA gets a higher value in an unseen music, then this music would give the arousal emotion that covers Excitement and Distress. We observed that the tonal feature is highly related to recognizing Depression and Contentment, because the TM2F and TM2S are the same tonal type features. It is also interesting to note that the Depression and Contentment consist of negative-arousal (sleepiness) in the emotion model. Thus, from the results, we can easily derive that the tonal type features are important to recognize the sleepiness of unseen music. This observation that the tonal feature is related to sleepiness is agreed on by human experts; they say that the tonal feature determines the softness of music. If the given music is a lullaby, then the tonal feature of this music would be soft.

Table 2. Top-ranked associative relations between selected musical features and emotion. The full name of feature is available from the website.

Emotion	A/V	Selected Features
Excitement	+/+	K1SS (Kurt.1 Spec.Kurt. Slope) D2DE (ddmfcc2 Delta-MFCC Ent.)
Distress	+/-	D1DA (dmfcc1 Delta-MFCC Amp.) I2SA (Irr.2 Spec. PeakAmp.)
Depression	-/-	TM2F (mode2 Keystrength Freq.) D1DA (dmfcc1 Delta-MFCC Amp)
Contentment	-/+	TM2S (mode2 Keystrength Slope) D1DE (dmfcc1 Delta-MFCC Ent.)

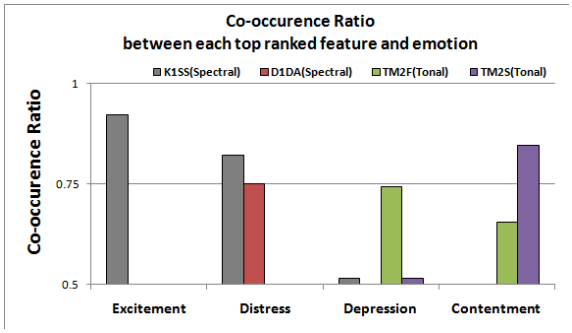


Fig. 4 Co-occurence ratio(>0.5) between top ranked features and each emotion. The feature K1SS showed a strong correlation (0.90) to Excitement emotion.

5. Conclusion and Future Works

Music emotion recognition has been popularized in music information retrieval area. Because music emotion recognition is highly depend on the relevant musical features, most of researcher select these features manually and it is definitely not scalable to massive musical features. In this paper, we presented a novel system to solve music emotion recognition problem. Our proposed system select features those are mutually far away from them while still being relevant to emotions. The proposed system showed better classification results and efficiency for the given data sets, indicating the potential of the proposed system.

Future work of this paper is as follow. Although we proposed a systematic approach to solve the redundancy among musical features, some part of our system still have

some opportunity to extend. For example, to recognize multiple music emotion, we employed parallel naive bayes, however, this method was originally proposed from traditional document categorization domain. To improve recognition accuracy, appropriate recognizer must be proposed.

References

[1] Y. Yang, Y. Lin, and H. Chen, "A regression Approach to Music Emotion Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 448-457, Feb, 2008.

[2] T. Li and M. Ogihara, "Content-based music similarity search and emotion detection," *In Proceedings of International Conference on Acoustic, Speech, Signal Processing*, pp. 17-21, Toulouse, France, 2006.

[3] L. Lu, D. Liu, H. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Transactions on Audio, Speech, and Language Processing.* vol. 14, no. 1, pp. 5-18, Jan, 2006.

[4] A. Hanjalic and L. Xu, "Affective video content representation and modeling," *IEEE Transactions on Multimedia*. vol. 7, no. 1, pp. 143-154, Feb, 2005.

[5] M. Korhonen, D. Clausi, and M. Jernigan, "Modeling emotional content of music using system identification," *IEEE Transactions on Systems Man and Cybernetics*, vol. 36, no. 3, pp. 588-599, Jun, 2006.

[6] Y. Yang, C. Liu, and H. Chen, "Music emotion classification: A fuzzy approach," *In Proceedings of ACM Multimedia 2006 (MM'06)*, pp. 81-84, Santa Barbara, CA, USA, 2006.

[7] J. Russell, "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161-1178, 1980.

[8] C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker, "A Matlab Toolbox for Music Information Retrieval," *Data Analysis, Machine Learning and Applications Proceedings of the 31th Annual Conference of the Gesellschaft fur Klassifikation e.V.*, Albert-Ludwigs-Universitat, Freiburg, Mar, 2007.

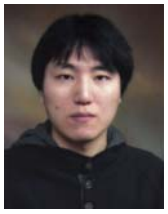
[9] Y. Feng, Y. Zhuang, and Y. Pan, "Music information retrieval by detecting mood via computational media aesthetics," *In Proceedings of IEEE/WIC International Conference on Web Intelligence*, pp. 235-241, Oct, 2003.

[10] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas, "Multilabel Classification of Music into Emotions," *In Proceedins of the 9th International Conference on Music Information Retrieval*, pp. 325-330, 2008.

[11] G. Tsoumakas, and I. Vlahavas, "Random K-labelsets: An ensemble method for multilabel classification," *In Proceedings of the 18th European Conference on Machine Learning (ECML 2007)*, pp. 406-417, Warsaw, Poland, Sep, 2007.

[12] H. Peng, F. Long, and C. Ding, "Feature Selection based on Mutual Information: Criteria of Max-dependency, Max-relevance, and Min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence.* vol. 27, no. 8, pp. 1226-1238, Aug, 2005.

[13] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining Multilabel Data [http://mulan.sourceforge.net/index.html]," Data Mining and Knowlege Discovery Handbook, O. Maimon, L. Rokach (Ed.), Springer, 2nd edition, 2010.



Jae-Sung Lee received the M.S. degree in Computer Science and Engineering, Chung-Ang University, Korea. He is currently in the Ph.D course. His research interests include data mining, classification, feature selection and affective computing.

Jin-Hyuk Jo is currently a M.S. candidate at Chung-Ang Univ. in Seoul, Korea, in the school of computer science and engineering, which he joined in 2009. He is currently interesting in data mining with applications to education data mining, specifically association rule mining and classification.

Jae-Joon Lee is research professor at Information Telecommunication Researching Institute, Chung-Ang Univ. He specializes in theory of digital art, pragmatist and phenomenological aesthetics of interaction, interdisciplinary researches between aesthetics and computer science, human-computer interaction theory and criticism, semantics of aesthetic data. Lee has taught contemporary aesthetics, the aesthetics of media, and digital art and culture theory in many universities since early 2000's.

Daewon Kim received M.S. and Ph.D degrees in computer science from Korea Advanced Institute of Science and Technology(KAIST), Daejon, Korea, in 1999 and 2004. Since 2005, he has been an assistant professor at the School of Computer Science and Engineering, Chung-Ang University, Korea. His research interests include data mining, pattern recognition, and artificial intelligence.