# Nearest Multi-Prototype Based Music Mood Classification

Babu Kaji Baniya, Choong Seon Hong
Department of Computer Engineering
Kyung Hee University
Yongin, Republic of Korea
babu775,cshong@khu.ac.kr

Joonwhoan Lee
Department of Computer Science and Engineering
Chonbuk National University
Jeonju, Republic of Korea
chlee@jbnu.ac.kr

*Abstract*—**Music mood classification is a crucial component in the field of multimedia database retrieval and computational musicology. There is a constantly growing interest in developing and evaluating music information retrieval (MIR) systems that can provide automated access to the music mood. The proposed method considers the different types of audio features. From each feature's frame, a bin histogram has been calculated to preserve all important information associated with it. The histogram bins of each feature are used to calculate the similarity matrix, and the number of similarity matrices depends on the number of audio features. Therefore, there are 59 similarity matrixes from the corresponding same amount of audio features. The intra and inter similarity matrix are used to calculate the intra-inter similarity ratio. These similarity ratios are sorted in descending order in each feature. Among them, some of the selected similarity ratios are ultimately used as prototypes from each feature and are used for classification by designing the nearest multi-prototype classifier. The Coimbra mood dataset is used to measure the overall performance of the proposed method. We achieved competitive classification accuracies as compared with other existing state-of-the-art music mood classification techniques.**

***Keywords—similarity matrix; multi-prototype; feature pool; histogram)***

## I. Introduction

Music can be viewed as categorical labels created by musicians or composers in order to find the content of the music. Music Information Retrieval (MIR) research activities are not only diversifying but also consistently growing. This diversity comes from the fact that emotion classification establishes certain relationships between music and its effect on human emotional states (e.g., happy, angry, sad, tender, fear, disgust etc.). Further, accessibility to music databases is increasing rapidly. As the amount of music content continues to increase, the search time is rapidly increasing. Having a database of music grouped under different emotions could lead to a reduction in the music information retrieval search time on online systems.

In this paper, similarity based music mood classification using a nearest multi-prototype classifier is shown in Fig. 1. It is an overview of our proposed method of mood classification. There are several problems that must be addressed in music mood classification, i.e. audio dataset collection, feature extraction, features selection from the feature pool, prototype selection, and a classifier design. For this purpose, a large number of audio features have been extracted. At first, we calculated the bin histogram of each feature. Therefore, the number of histograms depends on the number of corresponding features. In the next stage, a similarity measure is performed using dice similarity. The dice coefficient is the number of features common to both feature vectors relative to the average size of the total number of features present. The dice coefficient possesses intra and inter similarity, and so an intra-inter ratio can be calculated to design the prototypes of each feature. The prototypes play the deterministic role in which class of each song belongs to. So far, we choose the nearest-one prototype for class-label determination. Finally, music mood classification is performed by using the nearest multi-prototype classifier.

Features commonly exploited for music mood and genre classification can be roughly classified into timbral texture, rhythmic, harmony, or their combinations [1, 2]. Having extracted descriptive features, different pattern recognition algorithms are employed for their classification into mood. In this proposed method, a large number of audio features have been extracted, 59 in total, as shown in Table 1. They are divided into two categories, i.e., low and high-level, based on the frame size. The frame length for the low-level and high-level features were 46 ms and 2 s, respectively, with both having 50% overlap. Each frame level feature was used to make an eight bin histogram and was subsequently used to normalize it. Moreover, this approach makes it possible to preserve important frame level information obtained from the feature extraction.

The main contribution of this paper is to consider all extracted short time frames to make a histogram rather than to consider the integration parameters like mean, variance, skewness, kurtosis, covariance components etc. [3, 4, 5] of each song. Therefore, our method makes it possible to preserve the important frame level information of each song. Second, a similarity measure is performed by using dice similarity and an intra-inter ratio is also calculated to make prototypes. Third, a nearest multi-prototype classifier has been designed to find the mood label of each music piece to designate it as a particular class.

The outline of the paper is as follows. Feature extraction for automatic music mood classification is discussed in Section II. Section III describes the details of histogram calculation and

978-1-4799-8679-8/15/$31.00 copyright 2015 IEEE
ICIS 2015, June 28-July 1 2015, Las Vegas, USA

similarity measure procedures. Section IV describes prototype design and selection using intra-inter similarity ratio, while Section V briefly presents an experimental setup and dataset preparation used for music mood classification. Section VI provides the results and a discussion of similarity-based music mood classification accuracy, and is followed by a conclusion in Section VII.

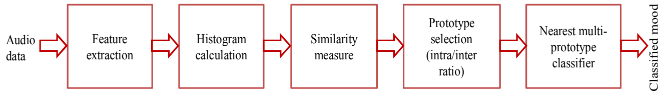


Fig.1. overview music mood classification using nearest multi-prototype classifier

## II. Feature Selection

Audio feature extraction encompasses the extraction of meaningful information from audio in order to obtain a compact and concise description that is machine-readable. Features are usually selected in the context of specific task and domain. Features are divided into two categories, i.e., low and high levels, according to the frame size. The frame length for the low-level and high-level features were 46ms and 2s, respectively, with both having 50% overlap. Because each frame-based feature provides a sequence of frame-wise values for audio data, this helps to determine a set of a small number of meaningful scalar values. MIR [6] and MA [7] toolboxes are used for feature extraction. The MA toolbox is essentially a subset of the MIR toolbox.

TABLE I. Audio Features for Music Mood Classification

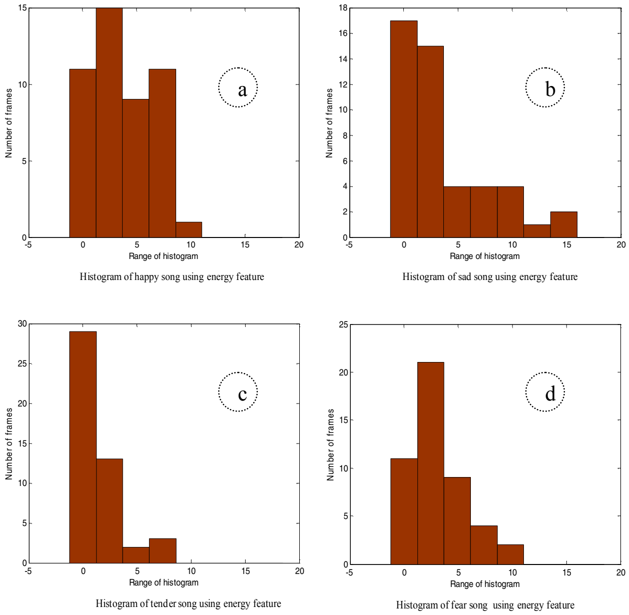
| No. | Category | Feature | Acronyms |
|---|---|---|---|
| 1 | Dynamic | RMS energy | Histogram |
| 2 | | Slope | ” |
| 3 | | Attack | ” |
| 4 | Rhythm | Tempo | ” |
| 5 | Spectral | Spectral centroid | ” |
| 6 | | Brightness | ” |
| 7 | | Spread | ” |
| 8 | | Rolloff85 | ” |
| 9 | | Rolloff95 | ” |
| 10 | | Spectral entropy | ” |
| 11 | | Flatness | ” |
| 12 | | Irregularity | ” |
| 13 | | Roughness | ” |
| 14 | | Zerocrossing | ” |
| 15 | | Spectral flux | ” |
| 16-28 | | MFCC(1-13) | ” |
| 29-41 | | DMFCC(1-13) | ” |
| 42-54 | | DDMFCC(1-13) | ” |
| 55 | Harmony | Chromagram peak | ” |
| 56 | | Chromagram centroid | ” |
| 57 | | Key clarity | ” |
| 58 | | Key mode | ” |
| 59 | | HCDF | ” |

## III. Histogram Calculation for Similarity Measure

At first, we tried to determine the minimum and maximum frame values of each audio feature from the whole dataset. These two values (minimum and maximum) are taken as reference points to calculate the 8 bin histogram of whole dataset for a particular feature. The aim of choosing the global reference points (i.e. minimum and maximum frame values) is to make clear distinctions between histograms. There are $H_1$, $H_2$ ,………, $H_n$ histograms from $F_1$, $F_2$, …, $F_n$ audio features where $n$ is the number of the feature. This histogram bin can be varied and can be used to evaluate the overall performance of a system according to histogram bins.

$$S_{dice} = \frac{2\sum_{i=1}^{d} \min(P_i Q_i)}{\sum_{i=1}^{d} P_i + \sum_{i=1}^{d} Q_i} \qquad (1)$$

where $P$ and $Q$ represent histograms of two different songs and $d$ is the number of bins in each histogram. The dice measure gives a square matrix of similarity values of features in each (mood) class. The similarity measure is divided into two categories, i.e., intra (within or inside the class) and inter (outside the class), respectively.



Histogram of happy song using energy feature

Histogram of sad song using energy feature

Histogram of tender song using energy feature

Histogram of fear song using energy feature

Fig.2. Histogram feature of music mood (a. fear, b. happy, c. sad and d. tender songs histogram) dataset using energy feature

## IV. Prototype Design and Selection

In the similarity matrix, intra similarity has single square matrix and inter has $N$-1 class size where $N$ be the number of class. The process of calculating of intra-inter similarity ratio algorithm is given below.

These intra-inter ratios have been sorted in descending order shown in Fig. 3. This can be done only training set of histogram features that are used to calculate the similarity matrix. The maximum intra-inter ratio object (index) has been taken first and finding the corresponding row of intra similarity matrix based on index number. In the Fig. 3, similarity ratio (highest value) of first index corresponds to the 14th row in intra similarity matrix. The first sorted value considered as a prototype. The similarity value of that intra similarity row subsequently compared with similarity ratio (inter-intra) table. If the similarity values of that row similar to any intra-inter ratio value than delete such object from the intra-inter ratio table. It is done by using some threshold. After comparing and deleting the similarity ratio based on intra similarity values in row 14th. We select another remaining highest value from intra-inter ratio table, find the corresponding index in intra similarity matrix, compare and delete object from intra-inter ratio table again. This is a second prototype. This process repeated until there is no intra-inter ratio remains in table except prototypes (sorted intra-inter ratio points out the corresponding to intra similarity row index). Once the intra-inter value is chosen as a prototype that doesn't take part in next comparison and deletion. This process continues till the total number of features in the proposed method. Based on these prototypes, we performed the validation by comparing the similarity with each feature's prototypes and take the majority voting to determine the class label of each song.

```
Class = NumberOfClasses
Feature = NumberOfFeatures
Sample_Num = ClassSize
Intra_Size = Sample_Num * Sample_Num // Size of intra class
Inter_Size = Sample_Num * Sample_Num*(Class-1) // Size of inter class
Intra_Inter = Sample_Num;
For i = 1:Feature
      Sum_Intra = 0.0
      Sum_Inter = 0.0
      For j = 1:Class
           Sum_Intra += Intra_Sim
           Sum_Inter += Inter_Sim
      End for$_j$
      Intra_Inter[i] = Sum_Intra[i]/ Sum_Inter[i]
End for$_i$
```

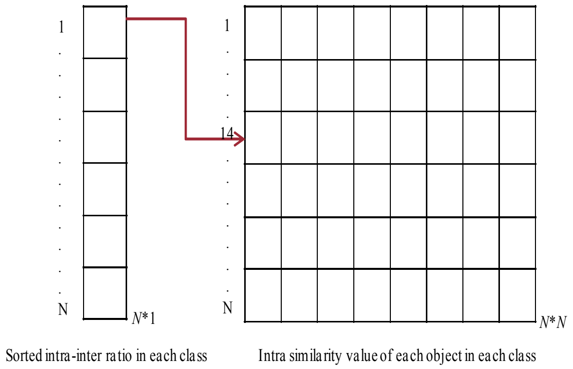

1
14
N
N*1
1
N
N*N
Sorted intra-inter ratio in each class
Intra similarity value of each object in each class

Fig.3. Sorted intra-inter index corresponds to the row of intra similarity matrix

## V. Experimental Setup and Data Preparation

We also took the Coimbra dataset for mood classification to verify whether our proposed method works well or not. This dataset is taken from the Informatics and System School of the University of Coimbra in Portugal. There are 903 mp3 audio files in total [8], which contain five different clusters with several emotional categories, given in Table 2.This dataset is almost balanced across clusters i.e., 18.8% cluster C1, 18.2% cluster C2, 23.8% cluster C3, 21.2% cluster C4 and 18.1% cluster C5.

TABLE II. Audio Features for Music Mood Classification

| Cluster | Adjectives | No. of songs |
|---|---|---|
| C1 | passionate, rousing, boisterous, rowdy | 170 |
| C2 | rollicking, cheerful, fun, sweet, amiable/good natured | 164 |
| C3 | literate, poignant, wistful, bittersweet, autumnal, brooding | 215 |
| C4 | humorous, silly, campy, quirky, whimsical, witty, wry | 191 |
| C5 | Aggressive, fiery, tense/anxious, intense, volatile, visceral | 163 |

## VI. Result and Discussion

A five-fold cross-validation was used to measure the performance of the proposed scheme (nearest multi-prototypes classifier) using Coimbra datasets. The overall classification accuracy of Coimbra dataset is 56.43% considering all audio features. The classification accuracy using audio, melody and combination of these features were reported 44.9%, 52.8% and 52.8%.0% features) respectively [8]. Similarly, they also performed the music mood classification using reduced features (both audio and melody features) and reported 64.0% classification accuracy in the same domain.

To get a better picture of the classification accuracy of the individual music mood, the confusion matrix is given in Table 3 Coimbra dataset. The confusion matrix is an $n$ x $n$ matrix, in which each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. Diagonal entries of the confusion matrix are the rates of music mood classification that are correctly classified, while the off-diagonal entries correspond to misclassification rates. The distribution of the Coimbra dataset is shown in Table 2, and the class level is represented by C1 to C5.

From the confusion matrix, it can be seen that Cluster 3 and 5 were classified with better accuracy than other clusters. Cluster 3 is mostly confused with Cluster 2 and 4. Similarly, the Cluster 5 is confused with 1 and 4. Cluster 2 and 4 had average classification accuracy. Likewise, Cluster 1 has the lowest classification accuracy among other clusters.

TABLE III. Confusion Matrix of Coimbra Dataset Using Different Feature Sets

| | C1 | C2 | C3 | C4 | C5 |
|---|---|---|---|---|---|
| C1 | **49.12** | 11.56 | 6.65 | 18.21 | 14.45 |
| C2 | 17.65 | **50.23** | 6.37 | 14.26 | 11.73 |
| C3 | 7.26 | 12.43 | **67.02** | 9.08 | 4.21 |
| C4 | 20.52 | 9.03 | 12.71 | **51.44** | 6.30 |
| C5 | 11.78 | 2.41 | 7.26 | 14.23 | **64.32** |

## VII. Conclusion

In this paper, we focus on designing the nearest multi-prototypes classifier for music mood classification. Before classifying the music mood, we collected diverse audio features belonging to low and high level features. From each feature, the eight bin histogram was calculated. The number of histograms depends on the number of extracted audio features in the proposed method and a similarity measure is performed by using dice similarity. The similarity matrix contains the intra and inter similarity in each class. Subsequently, the intra-inter ratio has been calculated using the similarity values, i.e., the sum of the intra similarity divided by the sum of the inter similarity. The similarity ratios are sorted in descending order in each feature. The first sorted value is considered as a prototype. Thereafter, the similarity value of that intra similarity row was compared with a similarity ratio (inter-intra) table. If the similarity values of that row are similar to any intra-inter ratio value, then the object is deleted from the intra-inter ratio table based on some threshold. After comparing and deleting the similarity ratios based on intra similarity values, a row is sorted using an index. Then, we select the remaining highest value from the intra-inter ratio table, find the corresponding index in the intra similarity matrix (the currently chosen intra-inter ratio index points to a row in the intra similarity matrix), and compare and delete the object in the intra-inter ratio table again. This is another prototype. Once the intra-inter value is chosen as a prototype, it does not take part in the next comparison (and deletion). This process continues to find prototypes until the intra-inter ratio remains in the ratio table. A similar process is used to find prototypes of other features too. Based on these prototypes, we performed validation by comparing the similarity with each feature's prototypes and take the majority voting to determine the class label of each song.

## Acknowledgment

This research was supported by Basic Science Research Program through National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2014R1A2A2A01005900)

## References

[1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals", *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Jul. 2002

[2] P. Saari, T. Eerola, and O. Lartillot, "Generalizability and Simplicity as Criteria in Feature Selection: Application to Mood Classification in Music", IEEE Trans. On Audio, Speech, and Lang. Process., vol. 19, no. 6, pp. 1802-12, Aug 2011.

[3] B. K. Baniya, D. Ghimire, and J. Lee, "Evaluation of different audio features for musical genre classification", *In proc. IEEE workshop on Signal Processing Systems*, Taipei, Taiwan, Oct. 2013

[4] B. K. Baniya, D. Ghimire, and J. Lee, "A Novel Approach of Automatic Music Genre Classification Based on Timbral Texture and Rhythmic Content Features", *Int. Conference on Advance Communication Technology (ICACT), pp.*96-102, 2014

[5] B. K. Baniya and J. Lee, "Importance of audio feature reduction in automatic music genre classification", Multimeida Tools and Applications, Dec. 2014

[6] E. Pampalk, "A Matlab toolbox to compute music similarity from audio", in *Proc. Int. Conf. Music Inf. Retrieval*, 2004 [Online]. Available: http://www.ofai.at/elias.pampalk/ma/

[7] O. Lartillot, "A Matlab toolbox to compute music similarity from audio", in *Proc. Int. Conf. Music Inf. Retrieval*, 2004 [Online]. Available: http://mir.dei.uc.pt/downloads.html

[8] R. Panda, R. Malheiro, B. Rocha, A. Oliveira, and R. P. Paiva, "Multi-Modal Music Emotion Recognition: A New Dataset, Methodology and Comparative Analysis", *International Symposium on Computer Music Multidisciplinary Research* 2013