Music Genre Classification Based on Timbral Texture and Rhythmic Content Features

Babu Kaji Baniya*, Deepak Ghimire**, Joonwhon Lee*
*Dept. of Computer Engineering, Chonbuk National University

Abstract

Music genre classification is an essential component for music information retrieval system. There are two important components to be considered for better genre classification, which are audio feature extraction and classifier. This paper incorporates two different kinds of features for genre classification, timbral texture and rhythmic content features. Timbral texture contains several spectral and Mel-frequency Cepstral Coefficient (MFCC) features. Before choosing a timbral feature we explore which feature contributes less significant role on genre discrimination. This facilitates the reduction of feature dimension. For the timbral features up to the 4-th order central moments and the covariance components of mutual features are considered to improve the overall classification result. For the rhythmic content the features extracted from beat histogram are selected. In the paper Extreme Learning Machine (ELM) with bagging is used as classifier for classifying the genres. Based on the proposed feature sets and classifier, experiment is performed with well-known datasets: GTZAN databases with ten different music genres, respectively. The proposed method acquires the better classification accuracy than the existing approaches.

1. Introduction

Automatic music genre classification is an important for information retrieval task since it can be applied for practical proposes such as organization of data collections efficiently in the digital music industry. There have been several well-known distinct approaches put forwarded on this. Still, efficient and accurate automatic music information processing stay at centre issue, and it has been consistently attracting the growing number of attention of researchers, musicians, and composers. A current challenging topic in automatic music information retrieval is the problem of organizing, describing, and categorizing music contents on the internet [1]. Although music genre classification is done manually, sometimes it is difficult to precisely define the genre of music content. The reason of difficulties is due to fact that music is state of art that evolves, where composers and musicians have been influenced by the music of other genre. Despite these difficulties, there are some possibilities still alive for genre classification. The audio signals of music belonging to the same genre mean that share the certain common characteristics, because they are composed of similar types of instruments, having similar rhythmic patterns, and similar pitch distributions [2]. The extracted features must be comprehensive (representing music very well), compact, and effective.

In this paper, we try to implement timbral texture features which represent short-time spectral information, and rhythmic content features like beat histogram which represent the long-term properties. Timbral texture features include spectral centroid, flux, rolloff, flatness, energy, zero crossing, MFCCs, respectively. We divide the timbral texture features into two groups for convenience; the first group (FG1) does not include MFCCs and the second group (SG2) includes only MFCCs. After the frame-wise extraction of each timbral texture feature among FG1 from all genres of music, next stage is to calculate the standard deviation for all genres of music. The aim of calculating the standard deviation of each feature in whole genres is to find out which feature is insignificant for genre discrimination. The feature which has small value of standard deviation contributes the insignificant impact on genre discrimination. Based on standard deviation value, we considered limited number of timbral features.

Similar procedure has been preceded for the SG2 of MFCC features as well. Out of thirteen, twelve coefficients give meaningful standard deviation values. This shows that twelve MFCC coefficients are meaningful for genre classification. For our experiment, we consider both first seven and twelve coefficients separately for genre classification.

In addition we propose to use the covariance components of mutual timbral texture features. Each of them gives the statistical property of mutual random variables associated features. For each song the covariance values of selected features from FG1 and SG2 calculated, respectively. Therefore, additional $n(n\text{-}1)/2$ components are included for n timbral texture features.

The feature set for representing rhythmic structure is based on detecting the most salient periodicities of the signal and it is usually extracted from beat histogram. Rhythmic content features contain relative amplitude of the first and second histogram peaks, period of first and second peaks, ratio of the amplitude of second peak divided by the amplitude of first peak, and overall sum of histogram.

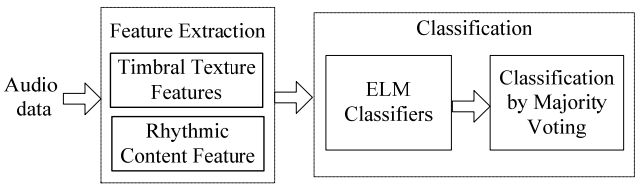

Fig.1. Overview of music genre classification

There are different types of classifiers which have been proposed for genre classification. We prefer distinct classifier

than previously applied one. Extreme Learning Machine (ELM) is recently proposed classifier which has high generalizing capability and takes minimum time for training.

2. Feature Extraction

Feature extraction encompasses the analysis and extraction of meaningful information from audio in order to obtain a compact and concise description that could be machine readable.

A. Timbral Texture features

These features are used to differentiate mixture of sounds that have possibly similar pitch and rhythm [3]. The features used to represent timbral texture are based on standard features proposed for music-speech discrimination [4]. To extract the timbral features, audio signals are first divided into frames by applying a windowing function at fixed intervals. The window function of this research is hamming window which helps to remove the edge effects. Timbral texture features have been computed and later we calculated different statistical values like mean, standard deviation, skewness, kurtosis, and covariance matrix from feature values. The mean (μ) and standard deviation (σ) for frame-wise feature values (X_n) in N-frame song are given by

$$Mean(\mu) = \frac{1}{N}\sum_{n=1}^{N} X_n \quad (1)$$

$$Std(\sigma) = \frac{1}{N}\sum_{n=1}^{N} (X_n - \mu)^2 \quad (2)$$

The skewness is a measure of asymmetry of the distribution, which can represent the relative disposition of the tonal and non-tonal components of each band. If the tonal components are frequently occurred in a band, the distribution of its spectrum will be left-skewed otherwise it will be right-skewed.

Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean. It is hard to specify the exact contribution of kurtosis in music genre classification [5].

Covariance is measured between two random variables or features. The aim of considering the covariance is usually to see if there is any relationship between the random variables. It is useful to measure the polarity and the degree of the correlation between two features.

1) FG1 features

Spectral flux: It is defined as the variation value of spectrum between the adjacent two frames in a short-time analyze window. It measures how quickly the power spectrum changes and is used to determine the timbral of an audio signal.

Spectral centroid: The spectral centroid is described as the gravity centre of the spectral energy. It determines the point in the spectrum where most of the energy is concentrated and is correlated with the dominant frequency of the signal.

Spectral flatness: It is used to characterize an audio spectrum. Spectral flatness is typically measure in decibels, and provides a way to quantify how tone like a sound is.

Spectral rolloff: It is a measure of the bandwidth of the audio signal. It is the fraction of bins in the power spectrum which 85% of the power is at lower frequencies.

Short Time Energy: The short time energy measurement of an audio signal can be used to determine voiced and unvoiced speech. It can also be used to detect the transition from unvoiced to voice and vice versa [6]. The energy of voiced speech is much greater than the energy of unvoiced speech.

Zero Crossing: It is a process of measuring the number of times in a given time interval that the amplitude of speech signals crosses through a value of zero. It is random in nature. Moreover, zero crossing rate for unvoiced speech is greater than that of voice speech.

2) FG2 features

Mel-Frequency Cepstral Coefficients:

Earlier MFCCs widely used in automatic speech recognition later on evolved into one of the prominent techniques in almost all domains in audio retrieval. They represent most distinctive information of signal. MFCCs have been successfully implemented to timbral measurements by H. Terasawa [7].

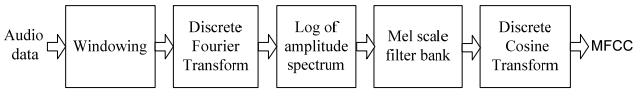

Fig. 2. Mel frequency cepstral coefficients feature extraction of audio.

A. Rhythmic Content Feature

Rhythmic content features characterize the movement of music signals over time and contain such information as the regularity of the rhythm, beat, and tempo. For rhythmic feature, beat histogram has been taken. It is a compact global representation of the rhythmic content of audio music. Beat histogram [8] can be obtained by wavelet decomposition of a signal and can be interpreted as successive high-pass and low-pass filtering of the time domain signal.

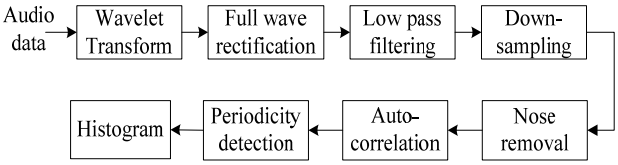

Fig.3. The block diagram of beat histogram for feature extraction

There are six different features extracted from beat histogram. They are relative amplitude of first and second histogram peak, period of first and second histogram peak measure in beat per minute (bpm), ratio of the amplitude of second peak divided by the amplitude of first peak, and average of overall bins.

3. Classifier

ELM [9] resolves the problem associated with gradient-based algorithm by analytically calculating the optimal weights of single-hidden layer feed-forward neural networks (SLFNs). Where the weights between input layers and the hidden layer biases are arbitrarily selected and then the optimal values for the weights between hidden layer and output layer are determined by calculating the linear matrix equations.

For N distinct samples and $\tilde{N}$ hidden nodes, the activation function g(x) of SLFN neural network defined as

$$\sum_{i=1}^{\tilde{N}} \beta_i g(w_i . x_j + b_i) = o_j, \ j = 1, ..., N \quad (19)$$

where $w_i = [w_{i1}, w_{i2}, .., w_{in}]^T$ is the weight vector connecting the ith hidden node and the input nodes, $\beta_i = [\beta_{i1}, \beta_{i2}, .., \beta_{im}]^T$ is the weight vector connecting the i-th hidden nodes and output nodes, and b_i is the threshold of the i-th hidden node. $w_i.x_j$ denotes the inner product of w_i and x_i.

The standard SLFNs with $\tilde{N}$ hidden nodes with activation function $g(x)$ can approximate these N samples with zero error means that $\sum_{j=1}^{\tilde{N}} \| o_j - t_j \| = 0$ i.e., there exist β_i, w_i, and b_i such that

$$\sum_{i=1}^{\tilde{N}} \beta_i g(w_i . x_j + b_i) = t_j, \ j = 1, ..., N \quad (20)$$

where t_j is the target vector of the j-th input data. Equation (19) can be written as a matrix equation to form new equation by using output matrix of the hidden layer H.

$$H\beta = T \quad (21)$$

where

$$H = \begin{bmatrix} g(w_1 . x_1 + b_1) & \cdots & g(w_{\tilde{N}} . x_1 + b_{\tilde{N}}) \\ \vdots & \cdots & \vdots \\ g(w_1 . x_N + b_1) & \cdots & g(w_{\tilde{N}} . x_N + b_{\tilde{N}}) \end{bmatrix}_{N \times \tilde{N}} \quad 22)$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_{\tilde{N}}^T \end{bmatrix}_{\tilde{N} \times m} \quad and \quad T = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}_{N \times m} \quad (23)$$

From the above equation (21), the target vector T and the output matrix of the hidden layer H can comprise a linear system. Thus, the learning procedure of the network helps to find the optimal weight matrix β between the output layer and the hidden layer β can be determined by using Moore-Penrose generalized inverse of H.

$$\hat{\beta} = H^{\dagger} T \quad (24)$$

Form the above equation (24) we can draw the following important properties. The first one is that we can take minimum training error, because the solution $\hat{\beta} = H^{\dagger} T$ is one of the least-square solutions of general linear system $H\beta = T$. In addition, the optimal $\tilde{\beta}$ is also minimum norm among these solutions. Thus, ELM has the best generalization performance than the typical back propagation network. In summary the ELM algorithm can be summarized as follows.

Algorithm ELM: For given training set $\aleph = \{(x_i, t_i) \mid x_i \in R^n, t_i \in R^m, i = 1, ..., N\}$, activation function $g(x)$, and hidden neuron number $\tilde{N}$,

1) Assign random input weight w_i and bias b_i, i=1,…, $\tilde{N}$.
2) Calculate the hidden layer output matrix H.
3) Calculate the output weigh β :

$$\hat{\beta} = H^{\dagger} T$$

Where $H^{\dagger}$ is the Moore-Penrose generalized inverse of hidden layer output matrix H.

4. Data preparation and result analysis

The dataset (GTZAN) consists of 1000 songs over ten different genres: Classical, Blues, Hiphop, Pop, Rock, Gazz, Reggae, Metal, Disco, and Country. Each class consists of 100 songs having 30s long duration. The dataset was collected by Geroge Tzanetakis [10]. Each song in database was stored as a 22050Hz, 16bits, and mono audio file. Five-fold cross validation scheme is used to evaluate the performance of the proposed system in GTZEN dataset.

TABLE I
CLASSIFICATION ACCURACY (CA) OF GTZAN DATASET IN DIFFERENT FEATURE SETS

Feature set	ELM(CA)
Energy+Centroid+Flux+Zerocrossing	68.33%
MFCC (7 coff)	64.62%
MFCC (12 coff)	66.26%
[ECFZ+MFCC, 7 coff.] [without covariance]+beat histogram	78.26%
[ECFZ+MFCC, 12 coff.] [without covariance]+beat histogram	80.21%
[ECFZ+MFCC , 7 coff.] [with covariance]+beat histogram	84.52%
[ECFZ+MFCC, 12 coff.] [with covariance]+beat histogram	85.58%

TABLE II
COMPARISON OF CLASSIFICATION ACCURACY WITH OTHER APPROACH OF GTZAN DATASETS (OUR APPROACH BASED ON FIVE-FOLD CROSS VALIDATION

Reference	CA
Our approach (ECFZ+MFCC+beat histogram)	85.58%
Jin S. Seo [11]	84.09%
Bergstra et al [9]	82.50%
Li et al.[1]	78.50%
Tzanetakis [5]	61.00%

First experiment was conducted within timbral texture features in FG1 like spectral centroid, flux, energy, and zero crossing for genre classification (excluding MFCC). The second and third experiments were only conducted for SG2 with seven and twelve mel-frequency cepstral coefficients (feature dimension shown in table II and III respectively). Fourth and fifth experiments only considered mean, standard deviation, skewness, and kurtosis of timbral texture feature including seven and twelve MFCC coefficients separately with rhythmic content feature. Final experiment was conducted taking covariance matrix and all other features. The experiment was performed into two different steps to find the classification accuracy considering minimum (seven MFCC coefficients) and maximum (twelve MFCC coefficients) feature dimensions considering mean, standard deviation, skewness, and kurtosis.

The combination of different feature sets gives different

TABLE III
CONFUSION MATRIX OF GTZAN DATASETS CLASSIFICATION ACCURACY WITH TIMBRAL TEXTURE AND RHYTHMIC CONTENT FEATURES

	Cl	Bl	Hi	Po	Ro	Ja	Re	Me	Di	Co
Cl	**95.0**	3.67	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.33
Bl	0.0	**85.25**	5.25	0.0	0.0	0.0	2.48	4.52	2.50	0.0
Hi	0.0	3.24	**81.26**	2.06	7.39	0.0	0.0	0.0	8.03	0.0
Po	0.0	0.0	0.0	**96.82**	1.14	0.0	0.0	0.0	2.04	0.0
Ro	2.18	1. 15	0.0	1.39	**74. 92**	0.0	3.55	7.09	5.03	2. 69
Ja	5.21	4.0	0.0	0.0	2.52	**83.12**	0.0	0.0	0.0	5.15
Re	0.0	0.0	0.0	8.64	2.23	3.68	**85.45**	0.0	0.0	0.0
Me	0.0	0.0	0.0	3.89	0.0	5.13	0.0	**89.57**	2.28	1.15
Di	0.0	0.0	2.35	0.0	4.95	0.0	6.25	2.35	**83.90**	0.0
Co	0.0	6.12	0.0	4.19	0.0	0.0	0.0	9.18	0.0	**80.51**

classification accuracy. The feature extracted from FG1 of energy, centroid, flux, and zero crossing gives 68.33 % of accuracy. Similarly, SG2(MFCC feature sets) with seven and twelve coefficients give 64.62% and 66.26% accuracy respectively. This experimental result shows that seven or eleven coefficients of MFCC do not make big difference in genre classification. We also tried to find out overall impact of classification accuracy with covariance components. The classification accuracy of GTZAN dataset without covariance components comes around 78.26% (7-MFCCs) and 80.21% (12-MFCCs) respectively. Among them, maximum accuracy is obtained while combining the covariance components with other feature sets. The classification accuracy of GTZAN dataset increases from 80.21% to 85.58% while including covariance components.

5. Conclusion

In this paper, first we analysed the validity of timbral texture features. The validity criterion is determined by normalized standard deviation of each feature. Second stage, the frame-wise features have been integrated by using central moments including mean, standard deviation, skewness and kurtosis. Also, we propose the covariance components between timbral texture frame-wise features to be included for improving the classification performance. By considering these feature values, several experiments have been performed separately to analysis classification accuracy among the different feature sets.

The classification accuracy of GTZAN dataset is shown in table II. According to our proposed method, the classification accuracy of 85.58% is achieved in GTZAN datasets.

Acknowledgement

This work was partially supported by the National Research Foundation of Korea grant funded by the Korean government (2011-0022152).

6. References

[1] Tao Li, Mitsunori Ogihara, and Qi Li, "A comparative study on content-based music genre classification", Proceedings of the 26th Annual International *ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 282-289, Toronto, Canada, 2003

[2] Xu, N. C. Maddage, and X. Shao, "Automatic music classification and summarization," *IEEE Trans.Speech Aud. Processing*, vol. 13,no. 3, pp. 441–450, May 2005.

[3] L. Rabiner and B. Juang. Fundamentals of Speech Recognition. Prentice-Hall, NJ, 1993.

[4] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), 1997, pp.1331-1334.*

[5] Jin S. Seo, Seungjae Lee, Higher-order moments for musical genre classification, *Signal Processing*, vol. 91, Issue 8, pp. 2154-57, 2011

[6] T.Zhang and C.J.Kuo, "Audio Content Analysis for Online Audiovisual Data Segmentation and Classification," *IEEE Trans. Speech and Audio Processing,* Vo1.9, No.4, pp. 441-457, May 2000.

[7] H. Terasawa, M. Slaney, and J. Berger. "Perceptual distance in timbrals space". In Proceedings of Eleventh Meeting of the International Conference on Auditory Display, pages 61-68 Limerick, Ireland, July 2005.

[8] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process*, vol. 10, no. 3, pp. 293–302, Jul. 2002.

[9] Huang, G.B.; Zhu, Q.Y.; Siew, C.K. Extreme Learning Machine: Theory and Applications. *Neurocomputing* 2006, 70, 489-501.

[10] Marasys, "Data sets" http://marsysas.info/download/data