

A Novel Approach of Automatic Music Genre Classification based on Timbral Texture and Rhythmic Content Features

Babu Kaji Baniya*, Deepak Ghimire*, Joonwhoan Lee*

*Division of Computer Enigneering, Chonbuk National University, Jeonju, South Korea
everwith_7@jbnu.ac.kr, deep@jbnu.ac.kr, chlee@jbnu.ac.kr

Abstract—Music genre classification is an essential component for the 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 the Mel-frequency Cepstral Coefficient (MFCC) with other several spectral features. Before choosing a timbral feature we explore which feature plays an insignificant 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 the classifier for classifying the genres. Based on the proposed feature sets and classifier, experiment is performed with well-known datasets: GTZAN with ten different music genres. The proposed method acquires better classification accuracy compared to the existing methodologies.

Keywords— music genre, bagging, covariance, timbral texture, rhythmic content

I. INTRODUCTION

A music genre is a conventional category that identifies pieces of music as belonging to a shared tradition or set of conventions. It is to be distinguished from musical form and musical style. Automatic music genre classification is important for the music information retrieval (MIR) task since it can be applied for practical proposes such as efficient organization of data collections in the digital music industry. There have been several well-known approaches put forward on this research area. Still, efficient and accurate automatic music information processing remains as the key issue, and it has been consistently attracting the attention of a growing number 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 and it takes huge resources for precisely define the genre of music content. Sometimes, it may create confusions and difficulties for its preciseness among composes and musician because of their perception.

Despite these difficulties, there are still some possibilities that remain for genre classification. The audio signals of music belonging to the same genre mean they share the certain common characteristics, because they are composed of similar types of instruments, having similar rhythmic patterns, and similar pitch distributions [2].

The overview of our music genre classification is shown in Figure 1. It depicts the backbone of genre categorization. There are two major problems that need to be addressed in genre classification, i.e., feature extraction and classification. The first stage is to extract the meaningful and relevant features from audio that could sufficiently discriminate the music genre. The next stage is to classify the genre based on the extracted features. In our method the extreme learning machine (ELM) combined with bagging is used as a classifier. Several bags of the dataset are constructed and each bag is trained using individual ELMs. The final decision is made based on the majority voting score. ELM is an unstable classifier, therefore ELM combined with bagging increases the stability, as well as generalization performance of the classifier.

In this paper, we attempt 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, and MFCCs, respectively. For the sake of convenient, we divide the timbral texture features into two groups; the first group (FG1) does not include MFCCs and the second group (SG2) only MFCCs. After the frame-wise extraction of each timbral texture feature among FG1 from all genres of music, the next stage is to calculate the standard deviation for all genres of music. The aim of calculating the standard deviation for each feature in whole genres is to find out which feature is insignificant for genre discrimination. The feature which has a small value of standard deviation contributes the insignificant impact on genre discrimination. Based on the standard deviation value, we considered a limited number of timbral features.

Timbral texture features are based on short time (10ms) low-level spectral components that are integrated on long duration. The integration method is mean and standard

deviation. Beside this, high order moments such as skewness and kurtosis are also implemented [3]. The aim of considering the high order moment is that even if there are the same values of mean and standard deviation, the position of location (shape of skewness and kurtosis position) could be different because each feature cannot be modelled the Gaussian distribution.

In addition we propose to use the covariance components of mutual timbral texture features [3]. 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 are calculated, respectively. Therefore, additional $n(n-1)/2$ components are included for n timbral texture features. Note that we can have $4n + n(n-1)/2$ for n features, which increases rapidly as the number of features increases. This is the reason why we remove the relatively less important features by checking the corresponding variances.

Rhythmic features characterize the movement of music signals over time and contain such information as the regularity of the rhythm, beat, tempo, and time signature. The feature set for representing the rhythmic structure is based on detecting the most salient periodicities of the signal and it is usually extracted from the 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 the second peak divided by the amplitude of first peak, and overall sum of the histogram [4].

There are different types of classifiers which have been proposed for genre classification. We prefer the distinct classifier than the previously applied one. Extreme Learning Machine (ELM) is a recently proposed classifier which has high generalizing capability and takes minimum time for training. The reason for selecting ELM is that it does not require a tuning parameter, has the smallest training error, and is free from the local maxima problem. However, ELM is unstable because the weights connected with hidden units are randomly determined. Therefore, we combine ELM with bagging in order to increase the stability. Bagging is almost always more accurate than a single classifier.

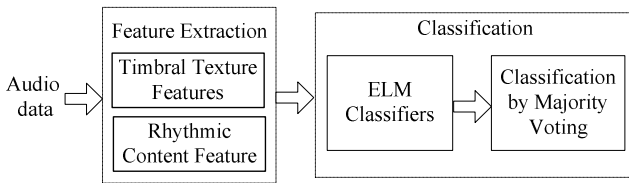


Figure 1. Overview of music genre classification

II. FEATURE EXTRACTION

This paper is organized as follows. Feature extraction is the critical portion of genre classification; and is described in section II. Section III deals with the classifier, similarly section IV explains the experimental setup and data preparation, and section V explains the result and analysis. Finally, section VI describes the conclusion of the proposed method.

A. Timbral Texture features

These features are used to differentiate mixture of sounds that possibly have similar pitch and rhythm [5]. The features used to represent timbral texture are based on standard features proposed for music-speech discrimination [6]. To extract the timbral features, audio signals are first divided into frames by applying a windowing function at fixed intervals. Timbral texture features in Figure 2 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 a N -frame song are given by

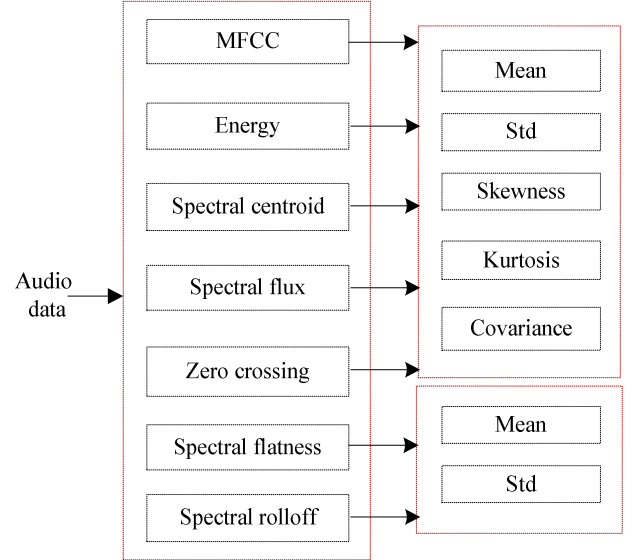


Figure 2. Overview of Timbral texture features extraction

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

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

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 occur frequently in a band, the distribution of its spectrum will be left-skewed otherwise it will be right-skewed. Mathematically, the skewness in a song can be defined as

$$\text{Skewness} = \frac{\sum_{n=1}^N (X_n - \mu)^3}{(N-1)\sigma^3} \quad (3)$$

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 difficult to specify the exact contribution of kurtosis in music genre classification [7]. However, the kurtosis measure can

sketch the effective dynamic range of the audio spectrum. Mathematically it can be defined as

$$Kurtosis = \frac{\sum_{n=1}^N (X_n - \mu)^4}{(N-1)\sigma^4} - 3 \quad (4)$$

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. The covariance of two features X_n and Y_n in a song is given as

$$Cov(X_n, Y_n) = \frac{1}{N} \sum_{n=1}^N (X_n - \mu_X)(Y_n - \mu_Y), \quad (5)$$

where μ_X and μ_Y are corresponding means of X_n and Y_n respectively. For n timbral texture features we acquire $n(n-1)/2$ mutual covariance values.

We consider two groups of timbral texture features FG1 and SG2 described as

1) FG1 features

Spectral flux: It is defined as the variation value of the 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.

$$F_t = \sum_{n=1}^N (N_t[n] - N_{t-1}[n])^2 \quad (6)$$

where $N_t[n]$ and $N_{t-1}[n]$ are normalized magnitudes of the Fourier transform at the present frame t , and previous frame $t-1$ respectively.

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. It is closely related to the brightness of a single tone.

$$C_t = \frac{\sum_{n=1}^N M_t[n] * n}{\sum_{n=1}^N M_t[n]} \quad (7)$$

where $M_t[n]$ is the magnitude of the Fourier transform at frame t and frequency bin n .

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 [8]. The energy of voiced speech is much greater than the energy of unvoiced speech. Short-time energy can be defined as

$$E_n = \sum_{m=1}^N [x(m)w(n-m)]^2 \quad (8)$$

where, $x(m)$ is discrete time audio signal, n is time index of short-time energy, and $w(m)$ is window of length N .

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, the zero crossing rate for unvoiced speech is greater than that of voice speech. Moreover, it is often used as a crucial parameter for voiced/unvoiced classification and end point detection.

$$Z_t = \frac{1}{2} \sum_{n=1}^N |\text{sgn}(x[n]) - \text{sgn}(x[n-1])| \quad (9)$$

where sgn is a short notation of sign function. The sgn function is 1 for positive arguments and 0 for negative arguments and $x[n]$ is the time domain for signal for frame t .

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

$$\sum_{n=1}^{R_t} M_t[n] = 0.85 \sum_{n=1}^N M_t[n] \quad (10)$$

where $M_t[n]$ is the magnitude of the Fourier transform at frame t and frequency bin n .

Spectral flatness: It is used to characterize an audio spectrum. Spectral flatness is typically measured in decibels, and provides a way to quantify how tone like a sound is, as opposed to being noise-like.

$$F = \frac{\exp\left(\frac{1}{N} \sum_{m=0}^{N-1} \ln x(m)\right)}{\frac{1}{N} \sum_{m=0}^{N-1} x(m)} \quad (11)$$

where $x(m)$ represents the magnitude of bin number m .

2) SG2 Features: Mel-Frequency Cepstral Coefficients
Earlier MFCCs widely used in automatic speech later on evolved into one of the prominent techniques in every domain of audio retrieval. They represent most distinctive information of signal. MFCCs have been successfully implemented to timbral measurements by H. Terasawa [9]. We considered the MFCC feature based on the paper that proved the validity of mel frequency cepstral coefficients for music modelling [10]. Figure 3 shows the process of obtaining MFCC features.

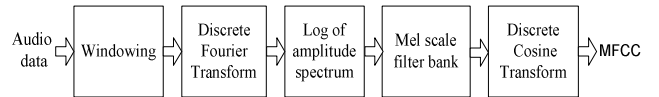


Figure 3. Mel frequency cepstral coefficients feature extraction of audio

B. Rhythmic Content Features

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 the rhythmic feature, beat histogram has been taken. It is a compact global representation of the rhythmic content of audio music. The beat histogram [4] can be obtained by the wavelet decomposition of a signal and can be interpreted as successive high-pass and low-pass filtering of the time domain signal.

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

the second peak divided by the amplitude of the first peak, and overall sum of the histogram. Feature extraction process is shown in Figure 4.

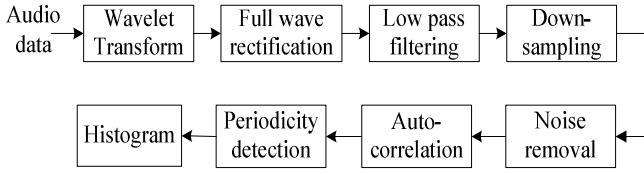


Figure 4. Block diagram of beat histogram feature extraction

III. CLASSIFIER

Traditionally, all the parameters of the feed-forward networks need to be tuned and thus there exists the dependency between different layers of parameters (weights and biases). In particular the gradient descent-based methods have been used in various learning algorithms of feed-forward neural networks [11]. However, the weakness of this kind of learning method is that it is generally very slow due to diverse learning steps and may easily converge to local minima. They also require many iterative learning steps in order to obtain better learning performance.

ELM [12] resolves the problem associated with the 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 the 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 the SLFN neural network is defined as

$$\sum_{i=1}^{\tilde{N}} \beta_i g(w_i \cdot x_j + b_i) = o_j, \quad j = 1, \dots, N \quad (10)$$

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

The standard SLFNs with $\tilde{N}$ hidden nodes with the 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

$$\sum_{i=1}^{\tilde{N}} \beta_i g(w_i \cdot x_j + b_i) = t_j, \quad j = 1, \dots, N \quad (11)$$

where t_j is the target vector of the j -th input data. Equation (10) can be written as a matrix equation to form a new equation by using the output matrix of the hidden layer H .

$$H\beta = T \quad (12)$$

where

$$H = \begin{bmatrix} g(w_1 \cdot x_1 + b_1) & \dots & g(w_{\tilde{N}} \cdot x_1 + b_{\tilde{N}}) \\ \vdots & \dots & \vdots \\ g(w_1 \cdot x_N + b_1) & \dots & g(w_{\tilde{N}} \cdot x_N + b_{\tilde{N}}) \end{bmatrix}_{N \times \tilde{N}} \quad (13)$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_{\tilde{N}}^T \end{bmatrix}_{\tilde{N} \times m} \quad \text{and} \quad T = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}_{N \times m} \quad (14)$$

From the above equation (12), 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 the Moore-Penrose generalized inverse of H .

$$\hat{\beta} = H^\dagger T \quad (15)$$

From the above equation (15) 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 the general linear system $H\beta = T$. In addition, the optimal $\hat{\beta}$ is also minimum norm among these solutions. Thus, ELM has the best generalization performance compared to the typical back propagation network. In summary the ELM algorithm can be summarized as follows.

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

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

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

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

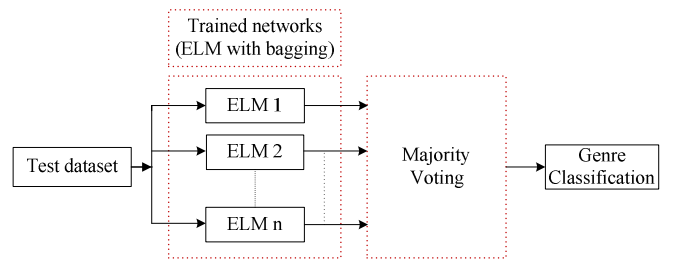


Figure 5: Block diagram of music genre classification

A. Bagging Algorithm

Bagging [13] is a well-known ensemble learning algorithm that has been shown to be very effective in improving generalization performance compared to the individual base models. Breiman indicated that bagging is a smoothing

operation which turns out to be advantageous when aiming to improve the predicative performance of regression or classification. It is a “bootstrap” ensemble method that creates bags for its ensemble by training each classifier on a random redistribution of the training set. Each classifier's training set is generated by randomly drawing, with replacement; many of the original samples may be repeated in the resulting training set while others may be left out. Each bag classifier in the ensemble is generated with a different random sample of the training set. The algorithm then applies a base classifier to classify each bag. Bagging is almost always more accurate than a single classifier. Finally the decision is taken by majority voting of all the base classifier results. Figure 5 shows the overview of genre classification. Our base classifier is ELM.

The Bagging Algorithm

Inputs: Training set S , based classifier L , integer T (number of bootstrap sample)

for $i = 1$ to T {

S_i =bootstrap sample from S (independent and identical component sample with replacement)

$E_i=L(S_i)$ }

$E^*(x) = \arg \max_{y \in Y} \sum_{i: E_i(x)=y} 1$ (the most often predicated label y)

Output: Compound classifier E^*

IV. EXPERIMENTAL SETUP AND DATA PREPARATION

A well known dataset widely used for music genre classification is employed for performance evaluation. The dataset (GTZAN) consists of 1000 music pieces divided into ten different genres: Classical, Blues, Hiphop, Pop, Rock, Gazz, Reggae, Metal, Disco, and Country. Each class consists of 100 music pieces having durations of 30s. The dataset was collected by Gerge Tzanetakis [14]. Each song in the database was stored as a 22050Hz, 16bits, and mono audio file. A five-fold cross validation scheme is used to evaluate the performance of the proposed system in GTZAN dataset.

V. RESULT AND ANALYSIS

At first, in order to reduce the dimensionality of the extracted feature set, the normalized standard deviation of each timbral texture feature is calculated in the GTZAN datasets. As the number of timbral texture feature increases the dimensionality of the extracted feature set increases rapidly, therefore we removed the relatively less important features by checking the corresponding normalized standard deviations. Table IV and V contain the average normalized standard deviation of the GTZAN datasets. The data seen in table IV indicates that flatness and rolloff are less significant for genre classification than the other four features in FG1. A similar approach is also applied for SG2 of MFCC coefficients. There is only one MFCC coefficient which has a relatively small value of normalized standard deviation. Aside from that coefficient, the remaining twelve coefficients are useful for genre classification.

The genre classification result of the GTZAN datasets is shown in table VI. Several experiments have been conducted among the different feature sets. The 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. The fourth and fifth experiments only considered mean, standard deviation, skewness, and kurtosis of timbral texture feature including seven and twelve MFCC coefficients separately with the rhythmic content feature. The final experiment was conducted taking the covariance matrix and all other features. The experiment was performed into two different steps to find the classification accuracy in regard to minimum (seven MFCC coefficients) and maximum (twelve MFCC coefficients) feature dimensions considering mean, standard deviation, skewness, and kurtosis.

TABLE IV: NORMALIZED STANDARD DEVIATION OF TIMBRAL TEXTURE FEATURES (EXCLUDING MFCC)

	Energy	Centroid	Flux	Zerocrossing	Flatness	Rolloff
Nor. std	1.085	0.690	0.886	0.671	0.239	0.274

TABLE V : NORMALIZED STANDARD DEVIATION OF MFCC

MFCC Coefficients	1	2	3	4	5	6	7
Normalized Standard Deviation	2.03	2.31	2.36	2.43	2.49	2.38	2.43
MFCC Coefficients (contd.)	8	9	10	11	12	13	
Normalized Standard Deviation	2.24	2.28	2.48	2.29	1.60	0.02	

TABLE VI: CLASSIFICATION ACCURACY (CA) OF GTZAN IN DIFFERENCE FEATURES

Feature set	Feature(D)	ELM(CA)
Energy + Centroid + Flux + Zerocrossing (ECFZ)	22	68.33%
MFCC (7 coff.)	49	64.62%
MFCC (12 coff.)	114	66.26%
ECFZ+MFCC (7 coff.)+beat histogram	77	84.52%
ECFZ+MFCC (12 coff.)+beat histogram	142	85.15%

The combination of different feature sets gives different 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 a big difference in genre classification. We also tried to find out the 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 the GTZAN dataset increases from

80.21% to 85.15% while including covariance components. Table VII compares our proposed method with other approaches in terms of average classification accuracy in the GTZAN dataset. It is clear that our proposed method achieves the classification accuracy of 85.15% which is better than other approaches.

TABLE VII: COMPARISON OF CLASSIFICATION ACCURACY (CA) WITH OTHER APPROACHES OF GTZAN DATASETS

Reference	Classification accuracy (CA)
Our approach	85.15%
Jin S. Seo [15]	84.09%
Bergstra et al [16]	82.50%
Li et al. [1]	78.50%
Tzanetakis [4]	61.00%

The extreme learning machine combined with bagging algorithm is used for the classification of the music genre. Before applied the ELM with bagging as a classifier, we attempted to find out how many bags were needed to obtain maximum classification accuracy. There were twenty three bags combined in the GTZAN dataset to get maximum classification as shown in Figure 6.

To get a better picture of the classification accuracy of an

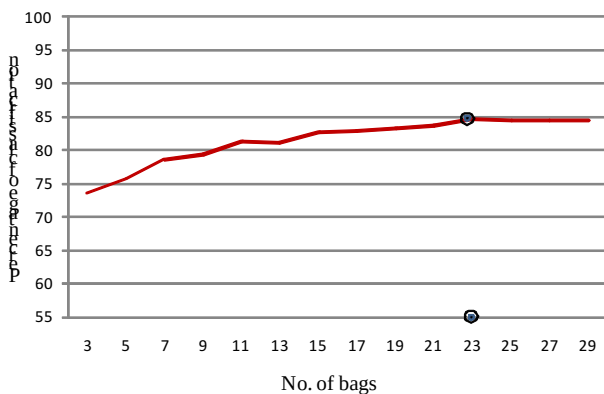


Figure 6: Number of bags vs classification accuracy of GTZAN dataset.

individual music genre, the confusion matrices are given. The confusion matrix is $n \times n$ matrix, at which each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The diagonal entries of the confusion matrix are the rates of music genre classification that are correctly classified, while the off-diagonal entries correspond to misclassification rates.

Table VIII shows the confusion matrix of the GTZAN dataset. The genres are arranged in the order of Classical (Cl), Blues (Bl), Hiphop (Hi), Pop (Po), Rock (Ro), Jazz (Ja), Reggae (Re), Metal (Me), Disco (Di), and Country (Co) respectively.

From the confusion matrix of the GTZAN dataset, we can see that some music genres are classified with significant

accuracy like Classical, Pop, Metal, and Reggae. Except for Rock, other music genre classification rates are also competitive. Rock music has a minimum classification rate. It is confused with Metal. Beside this, it is diverse in nature as compared to other genre and also overlaps its characteristic with other genres. Music genre Disco and Country are also confused with Reggae and Metal respectively.

VI. CONCLUSION

In this paper, first we analysed the validity of timbral texture features. The validity criterion is determined by the normalized standard deviation of each feature. In the 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 features to be included for improving the classification performance. By considering these feature values, several experiments have been performed separately to analyse classification accuracy among the different feature sets. The ELMs combined with bagging is used to build the classifier. The ELM is an unstable classifier; therefore ELMs with bagging improved the classification accuracy as well as the generalization performance.

Acknowledgement

This work was partially supported by the National Research Foundation of Korea grant funded by the Korean government (2011-0022152) and BK21PLUS project.

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] Babu Kaji Baniya, Deepak Ghimire, and Joonwhoan Lee, Evaluation of Different audio features for music genre classification, *Proceedings of the IEEE Workshop on Signal Processing Systems (SIPs), Taipei, Taipei*, pp 260-65, 2013
- [4] 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.
- [5] B. Logan. Mel frequency cepstral coefficients for music modeling. *In Proc. Int. Symposium on Music Information Retrieval ISMIR*, 2000.
- [6] Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, P. Cano, An Experimental Comparison of Audio Tempo Induction Algorithms, *IEEE Transactions on Speech and Audio Processing*, vol.14, page 1832-44, 2006.
- [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] 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.
- [9] 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.

- [10] Beth Logan, Mel Frequency Cepstral Coefficients for Music Modeling, *Int. Symposium on Music Information Retrieval* (2000)
- [11] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, pp. 251-257, 1991.
- [12] Huang, G.B.; Zhu, Q.Y.; Siew, C.K. Extreme Learning Machine: Theory and Applications. *Neurocomputing* 2006, 70, 489-501.
- [13] Beriman, L. Bagging Predictors. *Machine Learning* 1996, 24, 123-140.
- [14] Marasys, "Data sets" <http://marsyas.info/download/data>
- [15] Bergstra, J., Casagrande, N., Erhan, D., Eck, D. and Kegl B. "Aggregate features and AdaBoost for music classification", *Machine Learning*, Vol. 65, No. 2-3, pp. 473-484, 2006.
- [16] Jin S. Seo, Seungjae Lee, Higher-order moments for musical genre classification, *Signal Processing*, vol. 91, Issue 8, pp. 2154-57, 2011

BIOGRAPHIES



recognition etc.

Babu Kaji Baniya received the B.E. degree in Computer Engineering from Pokhara University, Nepal in 2005 and M.E. degree in Electronic Engineering from Chonbuk National University, Rep. of Korea in 2010. Currently he is pursuing his Ph.D. degree in Computer Science and Engineering at Chonbuk National University, Rep. of Korea from 2011. His main research interest includes audio signal processing, music information retrieval, source separation, pattern



Deepak Ghimire received the B.E. degree in Computer Engineering from Pokhara University, Nepal in 2007 and M.S. degree in Computer Science and Engineering from Chonbuk National University, Rep. of Korea in 2011. Currently he is pursuing his Ph.D. degree in Computer Science and Engineering at Chonbuk National University, Rep. of Korea from 2011. His main research interest includes image processing, computer vision, pattern recognition, facial emotion analysis etc.



include image processing, computer vision, and emotion engineering.

Joonwhoan Lee received the B.S. degree in electronic engineering from the Hanyan University, Seoul, Korea, in 1980, the M.S. degree in electrical and electronic engineering from KAIST, Yuseong, Korea, in 1982, and the Ph.D. degree in electrical and computer engineering from the University of Missouri in Columbia in 1990.

He is currently a Professor in the Department of Computer Engineering of the Chonbuk National University, Jeonju, Korea. His research interests