INTER GENRE SIMILARITY MODELLING FOR AUTOMATIC MUSIC GENRE CLASSIFICATION

Ulaş Bağcı, Engin Erzin

MVGL, College of Engineering
Koç University, Istanbul,Turkey
{ubagci|eerzin}@ku.edu.tr

ABSTRACT

Music genre classification is an essential tool for music information retrieval systems and it has been finding critical applications in various media platforms. Two important problems of the automatic music genre classification are feature extraction and classifier design. This paper investigates inter-genre similarity modelling (IGS) to improve the performance of automatic music genre classification. Inter-genre similarity information is extracted over the mis-classified feature population. Once the inter-genre similarity is modelled, elimination of the inter-genre similarity reduces the inter-genre confusion and improves the identification rates. Inter-genre similarity modelling is further improved with iterative IGS modelling(IIGS) and score modelling for IGS elimination(SMIGS). Experimental results with promising classification improvements are provided.

1. INTRODUCTION

Music genre classification is crucial for the categorization of bulky amount of music content. Automatic music genre classification finds important applications in professional media production, radio stations, audio-visual archive management, entertainment and recently appeared on the Internet. Although music genre classification is done mainly by hand and it is hard to precisely define the specific content of a music genre, it is generally agreed that audio signals of music belonging to the same genre contain certain common characteristics since they are composed of similar types of instruments and having similar rhythmic patterns. These common characteristics motivated recent research activities to improve automatic music genre classification [1, 2, 3, 4]. The problem is inherently challenging as the human identification rates after listening to 3sec samples are reported to be around 70% [5].

Feature extraction and classifier design are two pretentious problems of the automatic music genre classification. Timbral texture features representing short-time spectral information, rhythmic content features including beat and tempo, and pitch content features are investigated throughly in [1]. Another novel feature extraction method is proposed in [3], in which local and global information of music signals are captured by computation of histograms on their Daubechies wavelets coefficients. A comparison of human and automatic music genre classification is presented in a study [4]. Mel-frequency cepstral coefficients (MFCC) are also used for modelling and discrimination of music signals [1, 6].

Various classifiers are employed for automatic music genre recognition including K-Nearest Neighbor (KNN) and Gaussian Mixture Models (GMM) classifiers as in [1, 3], and Support Vector Machines (SVM) as in [3]. In a recent study, Boosting is used as a dimension reduction tool for audio classification [7].

In [8], we proposed *Boosting Classifiers* to improve the automatic music genre classification rates. In this study, in addition to inter-genre similarity modelling, two alternative classifier structures are proposed: i) Iterative inter-genre similarity modelling, and ii) Score modelling for inter-genre similarity elimination. Once the inter-genre similarity is modelled, elimination of those similarities from the decision process reduces the inter-genre confusion and improves the identification rates.

The organization of the paper includes a brief description of the feature extraction in Section 2. The discriminative music genre classification using the inter-genre similarity is discussed in Section 3. Later in Section 4 experimental results are provided following with discussions and conclusions in the last section.

2. FEATURE EXTRACTION

Timbral texture features, which are similar to the proposed feature representation in [1], are considered in this study to represent music genre types in the spectral sense. Short-time analysis over 25ms overlapping audio windows are performed for the extraction of timbral texture features for each 10ms frame. Hamming window of size 25ms is applied to the analysis audio segment to remove edge effects. The resulting timbral features from the analysis window are combined in a 17 dimensional vector including the first 13 MFCC coefficients, zero-crossing rate, spectral centroid, spectral roll-off and spectral flux.

3. MUSIC CLASSIFICATION USING INTER-GENRE SIMILARITY

The music signals belonging to the same genre contain certain common characteristics as they are composed of similar types of instruments with similar rhythmic patterns. These common characteristics are captured with statistical pattern recognition methods to achieve the automatic music genre classification [1, 2, 3, 4]. The music genre classification is challenging problem, especially when the decision window spans a short duration, such as a couple of seconds. One can also expect to observe similarities of spectral content and rhythmic patterns across different music genre types, and with a short decision window mis-classification and confusion rates increase. IGS modelling is proposed to decrease the level of confusion across similar music genre types. The IGS modelling forms clusters from the hard-to-classify samples to further eliminate the inter-genre similarity in the decision process of classification system. Inter-genre similarity modelling is further improved with iterative IGS modelling(IIGS) and score modelling for IGS elimination(SMIGS). IGS and its variations are described in the following subsections.

3.1. Inter-Genre Similarity Modelling

The timbral texture features represent the short-term spectral content of music signals. Since, music signals may include similar instruments and similar rhythmic patterns, no sharp boundaries between certain different genre types exist. The inter-genre similarity modelling (IGS) is proposed to capture the similar spectral contents among different genre types. Once the IGS clusters are statistically modelled with GMM, the IGS frames can be captured and removed from the decision process to reduce the inter-genre confusion.

Let $\lambda_1, \lambda_2, \ldots, \lambda_N$ be the N different genre models in the database($N = 9$ in our database). In this study, the Gaussian Mixture Models (GMM) are used for the class-conditional probability density function estimation, $p(f|\lambda)$, where f and λ are respectively the feature vector and the genre class model. The construction of the IGS clusters and the class-conditional statistical modelling(GMM) can be achieved with the following steps:

i. Perform the statistical modelling of each genre with GMM in the database using the available training data of the corresponding music genre class.

ii. Perform frame based genre identification task over the training data ,*(gmm test over training data)*, and label each frame as a true-classification or mis-classification.

iii. Construct the statistical model, λ_{IGS} with GMM, through the IGS cluster over all the mis-classified frames among all the music genre types.

iv. Update all the N-class music genre models, λ_n, using the true-classified frames.

The above construction creates N-class music genre models and a single-class IGS model. In the music genre identification process, given a sequence of features, $\{f_1, f_2, \ldots, f_K\}$, which are extracted from a window of music signal, one can find the most likely music genre class, λ^*, by maximizing the weighted joint class-conditional probability,

$$\lambda^* = \arg\max_{\lambda_n} \frac{1}{\sum_k \omega_{kn}} \sum_{k=1}^{K} \omega_{kn} \log p(f_k|\lambda_n) \qquad (1)$$

where the weights ω_{kn} are defined based on the class-conditional IGS model as following,

$$\omega_{kn} = \begin{cases} 1 & \text{if } p(f_k|\lambda_n) > p(f_k|\lambda_{IGS}), \\ 0 & \text{otherwise.} \end{cases} \qquad (2)$$

The proposed weighted joint class-conditional probability maximization eliminates the IGS frames for each music genre from the decision process. The inter-genre confusion decreases and the genre classification rate increases with the resulting discriminative decision process. Experimental results, which are supporting the discrimination based on the IGS elimination, are presented in Section 4.

3.2. Iterative Inter-Genre Similarity Modelling

Inter-genre similarity modelling can be repeatedly used to extend the detection of hard-to-classify samples in the training data. In each iteration a new IGS model is formed over the new set of mis-classified samples. In the decision process a frame is eliminated if it matches any one of the IGS classes. The construction of the T-step iterative inter-genre similarity (IIGS) models can be defined with the following steps:

i. Perform IGS modelling, and get λ_{IGS_1} and updated music genre models λ_n for all N-class. Set $t = 2$.

ii. Perform frame based genre identification task with IGS model over the training data, and label each frame as a true-classification, mis-classification or true-IGS classification.

iii. Construct the statistical model, λ_{IGS_t}, over all the mis-classified frames among all the music genre types.

iv. Update all the N-class music genre models, λ_n, over the true-classified frames.

v. Increment iteration counter t, and if $t \leq T$ go to step [ii.]

The above construction creates N-class music genre models and T-class IGS model. In the music genre identification process, the decision is taken by maximizing the weighted joint class-conditional probability in Equation 1. In IIGS modelling the weights ω_{kn} are re-defined based on the IGS models as following,

$$\omega_{kn} = \begin{cases} 1 & \text{if } p(f_k|\lambda_n) > p(f_k|\lambda_{IGS_t}) \text{for any } t = 1, ..., T \\ 0 & \text{otherwise.} \end{cases} \qquad (3)$$

3.3. Score Modelling for IGS Elimination (SMIGS)

In IGS modelling, the elimination of likely mis-classified frames is performed with a hard thresholding. That is, if IGS model produces the highest likelihood score for a frame, that frame is eliminated even though the best genre model is the true class and has a very close likelihood score to the IGS likelihood. Alternatively, IGS may score lower than the best likelihood score, while the best likelihood score belongs to a false-class. For such possible wrong decisions, the relative likelihood differences can be used as a reliability factor for the IGS elimination process. For example, if the likelihood difference between best two likelihood scores of music genre classes is low, one can claim that the decision is not reliable and the frame should be eliminated even though the IGS likelihood score is not the highest one. Hence, rather than taking a hard decision based on IGS likelihood score, one can model frame elimination based on the likelihood score distributions of the best M genre classes and the IGS class. Score modelling for IGS elimination is built on this idea, and described with the following procedure:

i. Perform IGS modelling, and get λ_{IGS} and updated music genre models λ_n for all N-class.

ii. Perform frame based genre identification task for each genre λ_n over the training data, extract the highest class conditional likelihood score that belongs to a class λ^* other than λ_n, $p(f_k|\lambda^*)$, $p(f_k|\lambda_n)$ and $p(f_k|\lambda_{IGS})$. Form a 3-dimensional likelihood score difference vector,

$$s_k = [\Delta_k^1 \ \Delta_k^2 \ \Delta_k^3]$$

where

$$\begin{aligned} \Delta_k^1 &= (p(f_k|\lambda_n) - p(f_k|\lambda^*)), \\ \Delta_k^2 &= (p(f_k|\lambda_n) - p(f_k|\lambda_{IGS})), \\ \Delta_k^3 &= (p(f_k|\lambda^*) - p(f_k|\lambda_{IGS})). \end{aligned}$$

iii. Construct the statistical models of the likelihood difference vectors for each genre λ_n over the true-classified and mis-classified frames of IGS modelling in step [i.], respectively as λ_{n_1} and λ_{n_0}.

The above construction creates N-class music genre models and for each genre true- and mis- classification models of the likelihood differences are extracted. In the music genre identification process, the decision is taken by maximizing the weighted joint class-conditional probability in Equation 1. In score modelling for IGS elimination the weights ω_{kn} are re-defined as following,

$$\omega_{kn} = \begin{cases} 1 & \text{if } p(f_k|\lambda_{n_1}) > p(f_k|\lambda_{n_0}) \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

4. EXPERIMENTAL RESULTS

Evaluation of the proposed classification algorithms is performed over a music genre database that includes 9 different genre types: classical, country, disco, hip-hop, jazz, metal, pop, reggae and rock. The songs in the database had been collected from CD collection of the authors and most of the songs are recorded from randomly chosen broadcast radios on the internet. The database includes totally 566 different representative audio segments of duration 30sec for all 9 music genre types, resulting a total duration of $566 \times 30 = 16980$ seconds. All the audio files are stored mono at 16000Hz with 16-bit words. The resulting timbral texture feature vectors are extracted for each 25msec audio frame with an overlapping window of size 10msec. The music genre classification is performed based on the maximization of the class-conditional probability density functions, which are modelled using the Gaussian mixture models (GMM). In experiments two-fold cross validation is used: the database is split into two partitions, each partition includes half of the audio segments from each genre type, where they are used in alternating order for training and testing of the music genre classifiers, and the average correct classification rates are reported in the following.

Decision	Correct Classification Rates (%)			
Window	Flat	IGS	SMIGS	IIGS
0.5s	44.61	49.95	52.70	50.46
1s	46.74	54.08	55.62	54.35
3s	49.22	58.50	61.99	59.16
30s	55.73	62.56	62.57	64.71

Table 1: *The average correct classification rates of the flat, IGS clustering, score modeling for IGS elimination (SMIGS) and iterative IGS clustering, using 8-mixture GMM modeling for varying decision window sizes.*

The three proposed music genre classification schemes, which are variations of inter-genre similarity elimination to reduce inter-genre confusion, are evaluated and compared with a flat classifier structure. Tables 1, 2 and 3 present correct classification rates of flat, IGS, SMIGS and IIGS classifiers for respectively 8, 16 and 32 mixture GMM modelling. Note that, IGS classification improves flat classification rates for all decision windows. We also observe further improvement using IIGS and SMIGS classifiers over IGS classifier. While the IIGS improvements are observed to be better for larger decision window sizes, the SMIGS improvements are better for smaller decision window sizes with 8-mixture GMM modeling. This behavior is expected, since iterative IGS expands the inter-genre similarity modelling, which causes elimination of more frame decisions, IIGS needs larger decision window sizes to bring further improvement.

Decision	Correct Classification Rates (%)			
Window	Flat	IGS	SMIGS	IIGS
0.5s	46.87	53.49	55.35	55.09
1s	49.32	56.80	57.06	58.71
3s	53.18	61.89	61.53	64.23
30s	57.78	66.91	67.26	72.48

Table 2: *The average correct classification rates of the flat, IGS clustering, score modelling for IGS elimination (SMIGS) and iterative IGS clustering, using 16-mixture GMM modelling for varying decision window sizes.*

However, with increasing number of GMM mixtures as presented in Tables 2 and 3, iterative IGS classifier is observed as the clear winner of these three discriminative music genre classification schemes at all decision window sizes. The classification rate improvements, which are achieved with IGS based classifiers, are significant, especially when the challenging automatic music genre classification task is considered with 70% human identification rate over 3s decision windows and 61% identification rate reported in [1].

Decision	Correct Classification Rates (%)			
Window	Flat	IGS	SMIGS	IIGS
0.5s	48.83	61.05	62.03	62.57
1s	51.81	65.28	66.64	66.96
3s	54.83	73.25	71.00	74.43
30s	59.90	83.16	81.58	84.41

Table 3: *The average correct classification rates of the flat, IGS clustering, score modelling for IGS elimination (SMIGS) and iterative IGS clustering, using 32-mixture GMM modelling for varying decision window sizes.*

5. CONCLUSION

Automatic music genre classification is an important tool for music information retrieval systems. Feature extraction and the classification design are the two significant problem of music genre classification systems. In feature extraction, a set of widely used timbral features are considered. In this work, we investigate three novel classifier structures for discriminative music genre classification. In proposed classifier structure, inter-genre similarities are captured and modelled over the mis-classified feature population for the elimination of the inter-genre confusion. The proposed iterative IGS model expands the inter-genre similarity modelling to better eliminate these similarities for the decision process. In score modelling for IGS elimination, a novel scheme based on statistical modelling of decision region of each genre for capturing IGS frames is presented. Experimental results with promising identification improvements are obtained with classifier design based on

the similarity measure among genres. Both in [8] and in this study we observed that IGS improves the flat classifiers and we also investigated the two extension of IGS modelling which increments the identification rates further depending on the number of mixture for GMM or size of the decision window.

6. REFERENCES

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

[2] D. Pye, "Content-based methods for managing electronic music," in *Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc. (ICASSP'00),* Istanbul, Turkey, 2000, pp. 2437 – 2440.

[3] T. Li, M. Ogihara, and Q. Li, "A comparative study on content-based music genre classification," in *Proc. 26th Int. ACM SIGIR conf. Research and Development in Information Retrieval*, 2003, pp. 282–289.

[4] S. Lippens, J. P. Martens, T. D. Mulder, and G. Tzanetakis, "A comparison of human and automatic musical genre classification," in *Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc. (ICASSP'04),* Montreal, Canada, vol. 4, May 2004, pp. 233–236.

[5] D. Perrot and R. Gjerdigen, "Scanning the dial: An exploration of factors in identification of musical style," in *Proc. Soc. Music Perception Cognition*, 1999, p. 88.

[6] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proc. Int. Symp. Music Information Retrieval (ISMIR'00),* Plymouth, Massachusetts, USA, 2000, pp. 138–147.

[7] S. Ravindran and D. Anderson, "Boosting as a dimensionality reduction tool for audio classification," in *Proc Int. Symp. Circuits and Systems (ISCAS '04)*, vol. 3, May 2004, pp. 465–468.

[8] U. Bağcı and E. Erzin, "Boosting classifiers for music genre classification," in *20th Int. Symp. on Computer and Information Sciences (ISCIS 2005), also appeared in Lecture Notes in Computer Science (LNCS) by Springer-Verlag*, Istanbul, Oct. 2005, pp. 575–584.