A Content-based Music Similarity Retrieval Scheme by Using BoW Representation and LSH-based Retrieval

Byeong-jun Han[1], Hyunwoo Kim[2], Ziwon Hyung[2], Kyogu Lee[2], Sheayun Lee[3]

[1] School of Electrical Engineering, Korea University, Seoul, Korea.
[2] Music and Audio Research Group (MARG),
Graduate School of Convergence Science and Technology (GSCST),
Seoul National University (SNU), Seoul, Korea.
[3] Bonacell, Co., Ltd., Seoul, Korea.

hbj1147@korea.ac.kr,
{kimbellw,ziotoss,kglee}@snu.ac.kr, sheayun@bonacell.co.kr

ABSTRACT

This extended abstract paper presents detailed information about a content-based music similarity retrieval scheme, which is based on locality sensitive hashing (LSH). Our scheme considered MFCC and time histogram (TH) as two major features to represent the properties of audio music similarity. Next, each feature is depicted by Bag of Words (BoW), which *k*-means clustering summarizes extracted features. In order to enhance the computational performance and preserve the result quality as much as possible, we adopted LSH to retrieve similar songs from dataset.

1. INTRODUCTION

Recent development and enhancement of mobility have made electronic consumers get more opportunities to experience music content via their own hand-held devices such as iPhone and Android OS-based phones. Especially, diverse interfaces on mobile devices, e.g. voice for singing and humming, tapping, drawing on touch screen, GPS, and gravity-sensitive sensors such as gyroscope, enables music service providers to imagine about music content service in diverse ways.

However, music analysis technology using music content instead of traditional textual information or user feedback needs significant improvement to provide music content without collecting user information. There are well-known audio descriptor standard like MPEG-7 audio descriptor[1] and its state-of-the-art implementation such as MARSYAS[2], nevertheless there are no killer implementations or services comparable with Last.fm[3].

In this paper, we introduce a content-based music similarity retrieval scheme which uses locality sensitive hashing (LSH) on bag of words (BoW) representation of music content information. Our scheme extracts two major features, the MFCC and time histogram (TH) in feature extraction step, and then depicts them into BoW by using *k*-means clustering algorithm. Finally, LSH is used to retrieve similar songs fast but not degrading the retrieval performance so much.

This paper is organized as follows. After this introduction section, section 2 describes the detailed information about our music similarity retrieval scheme by dividing into three steps: feature extraction; BoW representation, and LSH-based retrieval. Finally, section 3 summarizes overall content.

2. MUSIC SIMILARITY RETRIEVAL SCHEME

In this section, we present detailed information about proposed music similarity retrieval scheme. Our scheme consists of following three steps: feature extraction; BoW representation, and; LSH-based retrieval.

2.1 Feature Extraction

For content-based music analysis, feature extraction is important to bridge the gap between low-level representation and high-level semantic representation. In our scheme, we used MFCC and TH as fundamental features to represent the musical properties.

MFCC [4] is one well-known and being widely used feature in audio and speech analysis. The performance of timbre property representation have shown good result in many approaches, therefore, still many researchers are adopting MFCC as one of the most basic and fundamental features in audio/speech classification and detection tasks. However, one defect of MFCC is hardness to figure out the rhythmical or harmonic characteristics.

On the other hand, TH [5] depicts the rhythmical flow of overall song in moment. There are also diverse ways to extract tempo histogram information, e.g. novelty curve.

One defect on using MFCC solely or some relative features such as spectral features, e.g. spectral centroid, spread, flatness, kurtosis, is that the way does not reflect rhythm characteristic of song. Also, using TH only in music analysis task biases the result in the view point of rhythm characteristic. To get benefits from both timbre and rhythm features, we tried to integrate both features in next section.

2.2 Bag of Words (BoW) Model Representation

In order to simplify the extracted feature without degrading its uniqueness and quality, it is very important to select good method for BoW model representation [6]. In our method, we adopted well-known *k*-means clustering algorithm [7] for BoW representation method.

In our scheme, extracted features are clustered by *k* clusters. As the result, center of the clusters can be used as BoW for representing features into similar clusters.

One most important problem is lengths of the songs are different with each other. In order to obtain unique and characteristic but not affected from the length property of a song, our scheme generates a normalized histogram of words from a song, which counts the occurrence of words in representations.

On the other hand, our scheme also considers the transition by time flow. For example, transition from one word to another word by increasing one time step can be very useful information.

2.3 Locality Sensitive Hashing (LSH)

One critical issue in music similarity issue is, computational complexity to compare a song with other songs in database is very high. Usually in Breath-first search, it needs O(N^2), therefore it is hard to deal with under the scalable content providing environment.

LSH [8] can be one solution for this curse of the dimensionality problem. It is shown that LSH results good performance in many state-of-the art researches such as 3D object indexing[9], HTML style similarity[10], and human motion control[11].

LSH first generates limited number of observations, and then it observes and indexes the result into hash bins. By increasing the number of observations, the result becomes more accurate and similar with Breath-first search. Therefore, it is important to select the number of observations for finding the balance between computational complexity and quality of retrieved data.

3. CONCLUSION

This extended abstract paper introduces the details of our music similarity retrieval scheme, which consists of following three steps: MFCC and TH extraction for timbre and rhythm information acquisition; BoW representation by *k*-means clustering algorithm and statistical and transitional information representation, finally; LSH for solving the computational complexity problem in querying step with low degradation of retrieval performance.

4. REFERENCES

[1] MPEG-7 Part 4: Audio, ISO/IEC 15938:2002.

[2] G. Tzanetakis and P. Cook, “A framework for audio analysis,” *Organized Sound*, vol.4, no.3, 2000.

[3] Last.fm, http://last.fm .

[4] Meinard Meuller: *Information Retrieval for Music and Motion*, Springer, p.65, 2007.

[5] M. F. McKinney and D. Moelants, “Extracting the perceptual tempo from music,” *ISMIR 2004*.

[6] D. Lewis, “Naïve (Bayes) at forty: the independence assumption in information retrieval,” *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pp. 4-15, 1998.

[7] J.B. MacQueen, “Some methods for classification and analysis of multivariate observations,” *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281-297, 1967.

[8] P. Indyk and R. Motwani, “Approximate nearest neighbors: towards removing the curse of dimensionality,” *Proceedings of 30th Symposium on Theory of Computing*, 1998.

[9] B. Matei, *et al.*, “Rapid object indexing using locality sensitive hashing and joint 3D-signature space estimation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI)*, vol.28, no.7, pp.1111-1126, Jul. 2006.

[10] U. Tanguy, *et al.*, “Tracking web spam with HTML style similarities,” *ACM Transactions on the web (TWEB)*, vol.2, no.1, Feb. 2008.

[11] L. Ren, *et al.*, “Learning silhouette feature for control of human motion,” *ACM Transactions on Graphics (ToG)*, vol.24, no.4, Oct. 2005.