MIREX 2019: MULTI-SCALE CHROMA INDEXING

Jin S. Seo
Gangneung-Wonju National University
Dept. Electrical Eng.

ABSTRACT

This document describes our submission to 2019 MIREX cover song identification task based on the multiple-scale chromagram indexing. The feature extraction and the similarity measurement used for cover song search should cope with the wide range of possible distortions which may occur during cover song generation process. This paper deals with global and local tempo changes along with musical-key change by combining chroma n-gram, 2D FTM, and multi-scale indexing.

1. INTRODUCTION

The difficulty in cover song identification lies in the fact that there are lots of ways to generate cover versions of a song. Cover song identification method should cope with various modifications undergone during cover song generation while distinguishing one music signal from another. Among the modifications, key shift, tempo change, and signal displacement (or cropping) occur frequently. Another difficulty faced in deploying cover song identification systems, particularly those based on the local-similarity alignment, is the computational complexity and the time spent in retrieving potential cover versions. To handle massive amounts of music-related content over video-sharing services, query processing efficiency is important in practice. To deal with the mentioned issues, we propose a method which combines chroma n-grams, 2D Fourier transform magnitude (FTM), and multi-scale indexing in deriving song-level feature summary. The chroma n-grams, which are composed of n consecutive frame features, are extracted in multiple scales by successively resampling the input chroma sequence as a counter measure for tempo change and signal cropping. For query efficiency, we apply the k-means clustering over the extracted chroma n-grams and use the cluster centers as a song-level chroma feature summary. To attain key-shift invariance, we take 2D FTM of the cluster centers, which is termed as the multi-scale chroma n-gram index (MCNI). The MCNIs of two songs can be compared with the Euclidean distance, which is computationally less burdensome than dynamic time warping or SW distance.

2. MULTI-SCALE CHROMA N-GRAM INDEXING

An overview of the proposed MCNI is shown in Fig. 1. An audio signal is split into overlapping segments (called frames). From each frame, we extract an M-dimensional chroma vector (in this paper M = 12). Assuming that there are N frames in a music clip, the set of chroma vectors from each frame is given by

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}. \tag{1}$$

The chroma vector of a frame do not contain enough information for cover song identification, n-gram, which is a concatenation of the n consecutive frame features, has been used [3] for music retrieval. The local chroma characteristics of the music signal is depicted in the n-gram, which is the subsequence of $\mathbf{X}$ given by

$$\mathbf{G}_i = \{\mathbf{x}_i, \mathbf{x}_{i+1}, \cdots, \mathbf{x}_{i+(n-1)}\} \quad \text{for } 1 \leq i \leq N-(n-1) \ . \tag{2}$$

The nM-dimensional vector g_i is obtained by concatenating the elements of the $\mathbf{G}_i$. The n-gram vector g_i is not robust against tempo changes, which occur frequently during cover song generation process. To cope with tempo changes, previous works utilize tempo-insensitive distance measures [8] [3], such as dynamic time warping or SW distance, or the beat-synchronous chromagram [1] [2] which is obtained by aligning chroma sequence with musical beats. The tempo-insensitive distance measures are computationally demanding, and tracking of musical beats is difficult and error-prone due to the ambiguity and subjectivity of beat sensing. As an alternative to the previous approaches, we adopt multi-scale indexing [5], which has been widely-used in stereo matching and object recognition in computer-vision research. As shown in Fig. 1, the n-gram vectors are extracted in multiple scales by successively downsampling the input chroma vector sequences at a lower sampling rate (coarser resolution). The n-grams are extracted at different scales to deal with both global and local tempo changes. After collecting the nM-dimensional n-gram vectors g_i independently at all scales, we apply k-means clustering on them and obtain k cluster centers, which is used as a fixed-length song-level feature for cover song search. Each cluster center (the nM-dimensional vector) is reshaped into M-by-n signal where we apply 2D Fourier transform and take the magnitude of it to attain key-shift invariance as in the previous works [1] [4]. The Fourier transform magnitude is reordered into a vector, which is termed as MCNI. As shown in Fig. 1, an

MCNI vector is extracted from each cluster, and we finally have k MCNI vectors for a music signal. Details of the proposed method will appear at [7].

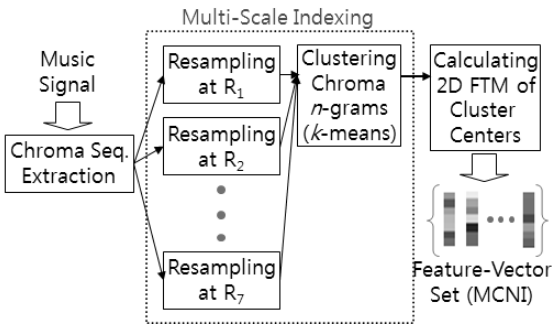

Figure 1. The extraction of the proposed multi-scale n-gram index from a music signal [7].

3. EVALUATION

The cover song search performance of the proposed multi-scale chroma n-gram indexing was evaluated on two cover song datasets. The first cover song dataset (abbreviated as covers80) is the one that was used by Dan Ellis in his work [2]. The covers80 consists of 80 original and cover song pairs, which are available online. For the covers80 dataset, we calculated the precision at one, P@1, which is the rate of the covers correctly identified in top 1 when querying each cover version on the 80 original songs.

For a fair comparison of retrieval performance, the same order chroma features were used for all the considered approaches ($M = 12$). Each song in the datasets was converted to mono at a sampling frequency of 22050 Hz and then divided into frames of 200 ms overlapped by 100 ms where the 12-dimensional chromagram was computed as a low-level feature for each frame. We extracted the chroma log pitch (CLP) using the chroma toolbox [6] with the default parameter settings. In deriving MCNI, we construct a scale-space representation by successively downsampling the extracted CLP with the seven scale levels (initial sampling rate: 2.6, final sampling rate: 1.5, and the scale factor between two levels of resolution: 0.913) as shown in Fig 1. The number of scales was set seven, which was the best compromise between the performance and the computation. For each song, the k-means clustering over the multi-scale n-gram vectors was computed 10 times, and the experimental results in this Section are the average performance of the 10 trials. We note that the standard deviation of the results from the 10 trials was small; the coefficient of variation (ratio of the standard deviation and the mean) was below 0.05 for P@1, which means that the experimental results are consistent regardless of the initialization of the k-means clustering used in deriving MCNI. The best P@1 of the proposed MCNI for covers80 dataset was 0.654 which is similar to that of the state-of-the-arts. The Euclidean distance associated with the proposed MCNI is computationally-efficient than the pairwise sequence alignment, which is practically important when dealing with large-size music repositories.

4. SUMMARY

In this paper, a multi-scale indexing method over chroma n-grams is proposed for cover song identification. The feature extraction and the similarity measurement used for cover song search should cope with the wide range of possible distortions which may occur during cover song generation process. This paper deals with global and local tempo changes along with musical-key change by combining chroma n-gram, 2D FTM, and multi-scale indexing. We derive index from the extracted n-grams by clustering to reduce storage and computation for DB search. Experimental results show that the multi-scale indexing is effective in improving cover song retrieval accuracy when combined with the conventional chroma n-gram and 2D FTM.

5. REFERENCES

[1] T. Bertin-Mahieux and D. Ellis. Large-scale cover song recognition using the 2D Fourier transform magnitude. In *Proceedings of ISMIR-2012*, pages 241–246, 2012.

[2] D. Ellis and G. Poliner. Identifying cover songs with chroma features and dynamic programming beat tracking. In *Proceedings of ICASSP-2007*, pages 1429–1432, 2007.

[3] P. Grosche and M. Muller. Toward characteristic audio shingles for efficient cross-version music retrieval. In *Proceedings of ICASSP-2012*, pages 473–476, 2012.

[4] J.H. Jensen, M.G. Christensen, and S.H. Jensen. A chroma-based tempo-insensitive distance measure for cover song identification using the 2D autocorrelation function. In *Proceedings of ISMIR-2008 (MIREX abstracts)*, 2008.

[5] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91 – 110, November 2004.

[6] M. Muller and S. Ewert. Chroma toolbox: MATLAB implementations for extracting variants of chroma-based audio features. In *Proceedings of ISMIR-2011*, pages 215–220, 2011.

[7] J.S. Seo. Multi-scale chroma n-gram indexing for cover song identification. *IEICE Transactions on Information and Systems*, 103-D(1), January 2020.

[8] J. Serra, E. Gomez, P. Herrera, and X. Serra. Chroma binary similarity and local alignment applied to cover song identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(6):1138 – 1151, August 2008.