LETTER

Special Section on Enriched Multimedia—Application of Multimedia Technology and Its Security—

Multi-Scale Chroma n-Gram Indexing for Cover Song Identification

Jin S. SEO†, Member

SUMMARY
To enhance cover song identification accuracy on a large-size music archive, a song-level feature summarization method is proposed by using multi-scale representation. The chroma n-grams are extracted in multiple scales to cope with both global and local tempo changes. We derive index from the extracted n-grams by clustering to reduce storage and computation for DB search. Experiments on the widely used music datasets confirmed that the proposed method achieves the state-of-the-art accuracy while reducing cost for cover song search.

key words: multi-scale indexing, cover song, music retrieval, n-gram, chromagram

1. Introduction

Considering millions of songs available in a large-size music archive nowadays, deriving a compact and discriminant music representation is utmost important for music retrieval. This paper deals with the cover song search among various music-retrieval tasks. Cover song refers to a live performance, a remix, or a new recording of a previously recorded track. The cover songs of the same musical piece might comprise large variations in timbre, rhythm, song structure, main key, and lyrics, which makes identifying a cover version of a song is interesting and challenging. The practical applications of the cover song identification are copyright protection and music-archive management.

For searching cover songs, auditory features representing the tonal contents of music, such as chromagram or pitch class profiles, have been utilized. The chromagram, which quantifies the spectral energy of the octave-folded subbands (in this paper, the frequency range of the twelve semitones) in a frame, is independent of timber and loudness and thus suitable for cover song identification [1]. Over the chromagram vector sequences, cover song search can be performed in two different ways; sequence alignment and song-level feature matching. Sequence-alignment based methods [1], [2] try to find the best alignment between feature sequences from two songs by adopting techniques used in speech recognition and DNA sequence identification, such as dynamic time warping or Smith-Waterman (SW) distance. By the way, the methods based on song-level feature matching [3] derive whole or segment-level feature summary from a song, which is then compared with the summary from another song to assess similarity between two songs. Both approaches have their own pros and cons. The sequence-alignment methods achieve better search accuracy but demand more computation and feature storage cost. By the way, the song-level feature matching can be deployed easily on commercial DB as shown in Fig. 1 using computationally-efficient distance measure, but has shown relatively poor accuracy due to the loss of local signal characteristics. This paper tries to improve the performance of the song-level feature summary by indexing chroma n-grams extracted in multiple scales.

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.

Fig. 1 Overview of the cover song search system with chromagram summary.
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.

The main contributions of our work are summarized as follows. 1) The proposed MCNI is robust against various common distortions incurred during cover song generation. 2) Computationally less burdensome distance measure can be incorporated with MCNI. 3) The proposed MCNI is compact by taking only cluster centers and thus can be easily implemented with a large-size music archive by reducing feature storage cost.

2. Proposed Cover Song Identification Method

In this section, we propose the MCNI by seamlessly combining the chroma $n$-gram and 2D chroma FTM with the multi-scale indexing to obtain robustness against key shift, tempo change, and cropping in deriving the song-level chroma representation.

2.1 Multi-Scale Chroma $n$-Gram Indexing

An overview of the proposed MCNI is shown in Fig. 2. 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

$$\mathbf{X} = \{ \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N \}. \quad (1)$$

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 [4] 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

$$\mathbf{g}_i = \{ \mathbf{x}_i, \mathbf{x}_{i+1}, \cdots, \mathbf{x}_{i+n-1} \} \quad (2)$$

for $1 \leq i \leq N - (n - 1)$.

The $nM$-dimensional vector $\mathbf{g}_i$ is obtained by concatenating the elements of the $\mathbf{G}_i$. The $n$-gram vector $\mathbf{g}_i$ 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 [1], [4], such as dynamic time warping or SW distance, or the beat-synchronous chromagram [3], [5] which is obtained by aligning chroma sequence with musical beats. The tempo-insensitive distance measures are computationally demanding, and tracking of musical beads 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 [6], [7], which has been widely-used in stereo matching and object recognition in computer-vision research. As shown in Fig. 2, 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 $\mathbf{g}_i$ 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 [3], [8]. The Fourier transform magnitude is reordered into a vector, which is termed as MCNI. As shown in Fig. 2, an MCNI vector is extracted from each cluster, and we finally have $k$ MCNI vectors for a music signal.

2.2 Musical Difference between MCNIs

In this paper, the musical difference between a query music clip and a song in DB is represented by the set difference between the two set of the MCNI vectors. The MCNIs of the query music and a song in DB are denoted by the two sets: $Q = \{ Q_1, Q_2, \cdots, Q_k \}$ and $A = \{ A_1, A_2, \cdots, A_k \}$ respectively. To obtain a set difference between $Q$ and $A$, we first calculate the $k$-by-$k$ pairwise Euclidean distance matrix $D_p$ given by

$$D_p[i, j] = \| Q_i - A_j \| \quad \text{for} \quad 1 \leq i, j \leq k \quad (3)$$

where $\| \cdot \|$ denotes the $L^2$ norm of the vector space. For each $Q_i$, we take the minimum distance to the set $A$ as follows:

$$d_{\text{min}}[i] = \min_{j} D_p[i, j]. \quad (4)$$

We sort the $k$-dimensional vector $d_{\text{min}}$ in ascending order to obtain the sorted distance $d_{\text{sort}}$. The set distance $D_{\text{set}}$ between $A$ and $Q$ is defined as the sum of the $T$ smallest distances by

$$D_{\text{set}}(Q, A) = \sum_{t=1}^{T} d_{\text{sort}}[t] \quad (5)$$
where \( T = \lfloor rk \rfloor \), and \( r \) is a control factor (typically \( r \) is between 0.7 and 1). For each query song, we calculate the set distances \( D_{set} \) with the songs in the DB and return the list of the closest songs as candidate cover versions.

3. Experimental Results

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 [5]. The covers80 consists of 80 original and cover song pairs, which are available online. The second cover song dataset (abbreviated as covers330) is composed of 1000 songs, where 330 songs are test data (30 original and corresponding 10 cover songs), and the other 670 songs were embedded as imposters. The covers330 was collected by the author. 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 the covers330 dataset, we queried the 330 cover songs over the 1000 entire songs in the dataset and evaluated the mean of average precision (MAP) following the same experimental procedures in MIREX 2018.

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 [9] 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. 2. 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 \( P@1 \) and 0.007 for MAP, which means that the experimental results are consistent regardless of the initialization of the \( k \)-means clustering used in deriving MCNI.

The retrieval accuracy of the proposed MCNI over different values of \( n \) and \( k \) is shown in Fig. 3 where we set \( r = 1 \) for calculating set distance in (5). As both \( n \) and \( k \) get larger, which needs more feature storage and computations, the cover song retrieval accuracy improves. However, for \( n \) larger than 60, we did not notice much performance improvement. With the value of \( k \) getting larger, the amount of the improvement in retrieval accuracy is getting smaller. In practice, on setting the value of \( n \) and \( k \), there is a trade-off between retrieval accuracy and system costs (i.e., feature storage and computations). The performance of the proposed MCNI over different values of \( k \) and \( r \) is shown in Fig. 4 where we use \( n = 65 \). Rather than using \( r = 1 \), which utilizes all \( k \) MCNI vectors in calculating set distance in (5), the value of \( r \) slightly smaller than one (typically between 0.7 and 0.9) was conducive in improving retrieval accuracy. The best \( P@1 \) of the proposed MCNI for covers80 dataset was 0.654, and the best MAP of the proposed MCNI for covers330 dataset was 0.706. We compared the performance of the proposed method with that of the OTI-SW [1]. The proposed MCNI yielded similar or better accuracy for both datasets; the \( P@1 \) of the OTI-SW for covers80 dataset was 0.55, and the MAP of the OTI-SW for covers330 dataset was 0.688. The set distance in (5) associated with the proposed MCNI is computationally-efficient than the pairwise sequence alignment of the OTI-SW, which is practically important when dealing with large-size music repositories. We note that the OTI-SW achieved one of the best performance in the MIREX audio cover song identification task [1]. Ex-

Fig. 3 (a) The precision at one, \( P@1 \), versus \( n \) for different values of \( k \) for covers80 dataset. (b) MAP versus \( n \) for different values of \( k \) for covers330 dataset.

\footnote{\textsuperscript{1}Accessed on October 2018 at http://www.music-ir.org/mirex/wiki/2018:Audio_Cover_Song_Identification}
experimental results show that the use of multi-scale indexing instead of the beat-aligned chromagram or sophisticated sequence matching can boost cover song retrieval accuracy.

4. Conclusion

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.

References