Audio Cover Song Identification using Convolutional Neural Network

Sungkyun Chang[1,4]**, Juheon Lee**[2,4]**, Sang Keun Choe**[3,4] **and Kyogu Lee**[1,4]
Music and Audio Research Group[1],
College of Liberal Studies[2],
Dept. of Electrical and Computer Engineering,[3]
Center for Superintelligence[4],
Seoul National University
{rayno1, juheon2, hana9000, kglee}@snu.ac.kr

Abstract

In this paper, we propose a new approach to cover song identification using a CNN (convolutional neural network). Most previous studies extract the feature vectors that characterize the cover song relation from a pair of songs and used it to compute the (dis)similarity between the two songs. Based on the observation that there is a meaningful pattern between cover songs and that this can be learned, we have reformulated the cover song identification problem in a machine learning framework. To do this, we first build the CNN using as an input a cross-similarity matrix generated from a pair of songs. We then construct the data set composed of cover song pairs and non-cover song pairs, which are used as positive and negative training samples, respectively. The trained CNN outputs the probability of being in the cover song relation given a cross-similarity matrix generated from any two pieces of music and identifies the cover song by ranking on the probability. Experimental results show that the proposed algorithm achieves performance better than or comparable to the state-of-the-art.

1 Introduction

In popular music, a cover song or cover version is defined as a new recording produced by someone who is not an original composer or singer. Cover songs share key musical elements, such as melody contours, basic harmonic progressions, and lyrics, with the original song. However, they can differ from the original song in other aspects, such as instrumentation, tempo, rhythm, key, harmonization, and arrangement. Applications of cover song identification include content-based music recommendation, detection of music plagiarism, and music sampling, to name a few.

Conventional methods for cover song identification generally combines a feature extraction and a distance metric. For feature extraction, chroma feature (Serra et al. [2009]) and its variants (Müller and Kurth [2006], Müller and Ewert [2010]) have been widely used for characterizing melodies and harmonic progressions. A distance metric then measures the similarity of sub-sequences in the feature space within two pieces of music. Various distance metrics, including dynamic time warping (DTW; Serra et al. [2008a]) cost, cross-correlation (Ellis and Cotton [2007]), and recently, similarity matrix profile (SimPLe; Silva et al. [2016]) and structural similarity (Cai et al. [2017])-based methods, have been proposed for this purpose.

So far, there have been a few attempts to exploit machine learning for cover song identification. Humphrey et al. [2013] used sparse coding with 2-dimensional Fourier magnitude coefficient derived from chroma. Recently, Heo et al. [2017] attempted to apply metric learning (Davis et al. [2007]) to results from SimPLe. Both of these works were based on existing deterministic cover song
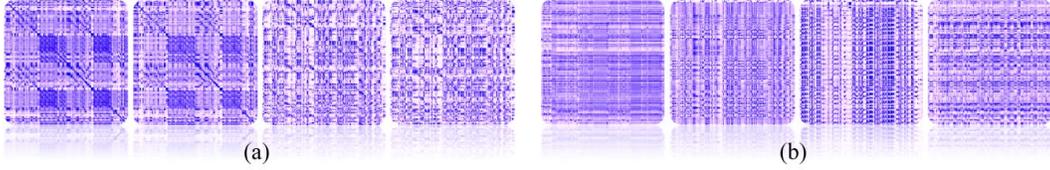
Figure 1: Sampling of 180×180 cross-similarity matrices generated using the first 180 s of each song. (a) were generated from the cover pairs of one original song "Toy - Passionate goodbye" and its four different cover versions. (b) were generated using the same original song and its four different non-cover pair.

identification algorithms, and they mainly focused on improving the scalability of cover song discovery by proposing a novel embedding technique or metric subspace learning for the distance calculation, respectively.

In this research, we propose a convolutional neural network-based system for audio cover song identification. We use a cross-similarity matrix generated from a pair of songs as an input feature. This idea is based on the observation that similar sub-sequences within cover songs often appear as a meaningful pattern in the cross-similarity matrix. With this assumption, we reformulate the audio cover song identification problem in the image classification framework.

2   Basic Idea

In various previous works on audio matching, the local chroma energy distributions across a shifting time window have been widely used as a representation of pitch contents, including melody contour and chord progression. Based on Hu et al. [2003], we first convert the audio signals for each song into a 12-dimensional chroma feature with a 1 s non-overlapping window. Then, we can define a cross-similarity matrix S with respect to a pair of two chroma features $\{A, B\}$ as

$$S_{l,m} = \frac{\max(\Delta) - \Delta_{l,m}}{\max(\Delta)}, \quad \text{s.t.} \ \Delta_{\{l,m|l \in L, m \in M\}} = \delta(A_{(:,l)}, B_{(:,m)}), \tag{1}$$

where δ denotes a distance function, and $\{L, M\}$ are the entire time indices of chroma sequence $\{A \in \mathbb{R}^{12 \times L}, B \in \mathbb{R}^{12 \times M}\}$, respectively. For δ, we calculate the Euclidean distance after applying the key alignment algorithm proposed in Serra et al. [2008b]. This is also known as the optimal transposition index.

Fig. 1 displays eight examples of the S generated from (a) the four cover pairs and (b) four non-cover pairs. The leftmost two images of (a) were generated from cover pairs containing almost same accompaniments, and we could observe consistent diagonal stripes with block patterns. In the third and fourth leftmost images of (a) were generated from the cover pairs produced in different tempo and instrumentations. Although the block patterns disappeared, we could observe consistent diagonal stripes in contrast with (b) from the non-cover pairs. Based on this observation, we assume that a convolutional neural network model for image classification can distinguish relevant patterns from the cross-similarity matrix. More specifically, a block of convolutional layers can sequentially perform sub-sampling and cross-correlation (or convolution) for distinguishing meaningful patterns from images in many different scales. Currently, we only compare the first 180 s of each song: We observed that most of popular music recordings had durations of three to five minutes, and the first three minutes mostly contains main melodies. Thus, we assume that the first 180 s of each song could provide relevant information to identify a cover song. If the song lasted for less than 180 s, the duration of the song was standardized with zero-padding.

Note that Eq. 1 is equivalent to the intermediate process of SimPLe proposed in Silva et al. [2016]. Another closely related work is Sakoe and Chiba [1978], which exploits a cross-similarity matrix in the early process of speech alignment. In addition, similar ideas of utilizing stripe or block-like patterns in a self-similarity matrix have been proposed in various works for audio music segmentation (Paulus et al. [2010]). All these findings motivated us to use the cross-similarity matrix with a convolutional neural network.

Input audio (pairs) → Preprocessing → Cross-similarity matrix → CNN → Softmax output → Ranking → Top@*N* covers
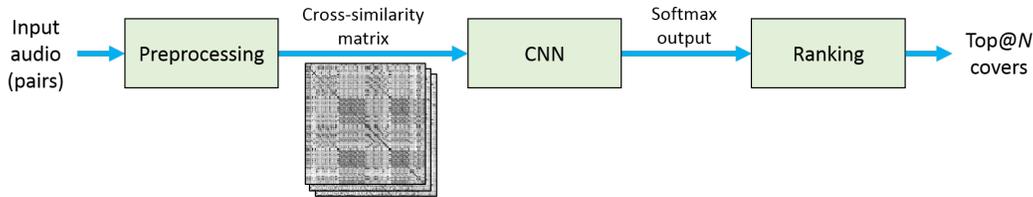
Figure 2: Overview of the proposed system.

Table 1: Specification of convolutional neural network: Inside the brackets are unit convolutional blocks, and outside the brackets is the number of stacked blocks. *Conv* denotes a *same* convolution layer with stride = 1, and its inside parentheses is (channel×width×height). *Maxpool* denotes a max-pooling layer with stride = 1, and its inside parentheses is (pooling size). *BN and FC* denote batch normalization and fully-connected layer, respectively.

Block #	Input layer	block 1		block 2	block 3	block 4	block 5	Final layers	
Components	-	*Conv* (32 × 5 ×5), ReLU *Conv* (32 × 5 ×5), ReLU *Maxpool* (2 × 2) *BN*	× 1	*Conv* (32 × 3 ×3), ReLU *Conv* (16 × 3 ×3), ReLU *Maxpool* (2 × 2) *BN*			× 4	*DropOut^p(0.5)* *FC(256)*, ReLU *DropOut^q(0.25)*	*FC(2)* softmax
Output shape	(1, 180, 180)	(32,90,90)		(16,45,45)	(16,22,22)	(16,11,11)	(16,5,5)	(,,256)	(,,2)

3 Proposed System

The proposed system, shown in Fig. 2, consists of three stages. In the preprocessing stage, we convert audio signals into chroma features for each song. Then we generate cross-similarity matrices by taking a pair of chroma features as described in Section 2.

The next stage is based on the convolutional neural network (hereafter, CNN), as specified in Table 1. Our CNN is built as a narrower and deeper network (0.58×10^6 parameters with 10 convolution layers) than conventional CNNs for ImageNet, such as AlexNet(Krizhevsky et al. [2012]) which has 60×10^6 parameters with five convolution layers. With respect to the size of the input cross-similarity matrix, we currently fix it as 180×180 (cut or zero-padded) that corresponds to comparing the first 3 min of music. With respect to the filter size of the first convolution layer, the receptive field of the first layer corresponds to 5 s of audio (2–4 measures in a music score). In practice, using the first convolution filter size of 5×5 resulted in approximately 4 % better performance than using 3×3 or 7×7. With respect to blocks 2–4, the basic idea in Section 2 was to run a chain of processing pattern consisting of sub-sampling and cross-correlation (or convolution) with these blocks. For this, blocks 2-4 of the CNN are built using a template convolutional block that outputs a one-half down-sampled size. In every convolutional block, we apply batch normalization (Ioffe and Szegedy [2015]).

The last stage of our system performs ranking on the softmax output of the trained CNN. We first take the cover-likelihood vector over all cover candidates. Then we apply descending-sort on this vector for ranking the most likely top@N covers.

4 Experimental Results

4.1 Data set

We use an evaluation data set provided by Heo et al. [2017]. The data set resembles that used for the MIREX[1] cover song identification task. It consists of 330 cover songs that make the query set, and 670 dummy songs that are not covered. Of the 330 query songs, there are 30 different kinds of cover songs. Each has 11 different cover versions (each query song must have 10 ground-truth covers). Thus, it can yield test examples of 3,300 cover pairs and 496,200 non-cover pairs.

The training set consists of 2,113 cover pairs and 2,113 non-cover pairs. The held-out validation set consists of 322 cover pairs and 322 non-cover pairs. These data sets are disjoint. The audio files

---

[1] http://www.music-ir.org/mirex/wiki/

contain popular Korean music released from 1980 to 2016. They were produced in stereo with a sampling rate of 44,100 Hz.

4.2 Training

In advance of the training, we applied zero-mean unit standardization on the input cross-similarity matrices for feature scaling. We trained the CNN with a total of 4,226 cross-similarity matrices (class-balanced for cover and non-cover) . The CNN was implemented based on the Keras framework, and run on a single GPU cloud server. Using the Adam optimizer (Kingma and Ba [2014]), the training stopped when the cross entropy loss ϵ reached convergence for $\epsilon < 10^{-4}$. Using a nested grid-search, we tried to optimize the two drop-out hyperparameters, denoted as drop-out^p and drop-out^q in Table 1. We achieved the final validation accuracy 83.4% for drop-out^p (0.5) and drop-out^q (0.5), by not looking at test set accuracy.

4.3 Results

We evaluated the proposed system by following metrics proposed in the MIREX for audio cover song identification task:

- MNIT10: mean number of covers identified in top 10.
- MAP: mean average precision.
- MR1: mean rank of the first correctly identified cover.

Here, MNIT10 was calculated as {total number of correctly identified covers in top 10} divided by {total number of ground-truth covers (= 3,300)}.

Table 2: Performance of audio cover song identification

Model	MNIT10	MAP	MR1
SimPLe Silva et al. [2016]	6.8	0.66	5.6
SimPLe + Metric Learning (Heo et al. [2017])	7.9	0.81	15.1
CNN (proposed)	8.04	0.84	2.50

In Table 2, we compared our system with two baseline algorithms: Silva et al. [2016], a deterministic algorithm, and Heo et al. [2017], a metric learning-based algorithm. The largest MNIT10 was achieved by the proposed CNN. This implies that the search result of the proposed system contained 8.04 correct covers out of 10, in average. With respect to MNIT10 and MAP (where larger is better), the present CNN showed competitive precision over the two compared algorithms. With respect to MR1 (where smaller is better), the proposed CNN achieved 80.10% improved performance over SimPLe, the second-best algorithm. The smaller MR1 implies that the ground-truth covers would more consistently appear in top search results. The effect of comparing various input lengths of each song has not been examined yet. However, the proposed system comparing only the first 180 s achieved the better performance over than all other systems comparing the entire lengths of input songs.

5 Conclusions and Future Work

We proposed a convolutional neural network-based approach to audio cover song identification. Our assumption was that the cross-similarity matrix from a pair of two songs could appear as a meaningful pattern. Based on this, we trained the CNN using cross-similarity matrices in the same manner that a binary classifier for images is trained. By ranking the softmax output from the trained CNN, the proposed system was able to predict a fixed number of the most likely cover song pairs. The performance of the proposed system was compared with a deterministic approach and another machine learning-based approach. Although the current study showed promising results, there is much room for improvement, particularly by finding more a suitable CNN design, hyper-parameter tuning, and increasing the size of the training data set with flexible input feature length. Furthermore, we did not apply any of the embedding techniques that are necessary for a large-scale search of cover songs. Thus, exploration of these is left for future work.

References

Joan Serra, Xavier Serra, and Ralph G Andrzejak. Cross recurrence quantification for cover song identification. *New Journal of Physics*, 11(9):093017, 2009.

Meinard Müller and Frank Kurth. Towards structural analysis of audio recordings in the presence of musical variations. *EURASIP Journal on Advances in Signal Processing*, 2007(1):089686, 2006.

Meinard Müller and Sebastian Ewert. Towards timbre-invariant audio features for harmony-based music. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):649–662, 2010.

Joan Serra, Emilia Gómez, Perfecto Herrera, and Xavier Serra. Chroma binary similarity and local alignment applied to cover song identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(6):1138–1151, 2008a.

Daniel PW Ellis and C Cotton. The 2007 labrosa cover song detection system. *MIREX extended abstract*, 2007.

Diego F Silva, Chin-Chin M Yeh, Gustavo Enrique de Almeida Prado Alves Batista, Eamonn Keogh, et al. Simple: assessing music similarity using subsequences joins. In *International Society for Music Information Retrieval Conference, XVII*. International Society for Music Information Retrieval-ISMIR, 2016.

Kang Cai, Deshun Yang, and Xiaoou Chen. Cross-similarity measurement of music sections: A framework for large-scale cover song identification. In *Proceeding of the Twelfth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Nov., 21-23, 2016, Kaohsiung, Taiwan, Volume 1*, pages 151–158. Springer, 2017.

Eric J Humphrey, Oriol Nieto, and Juan Pablo Bello. Data driven and discriminative projections for large-scale cover song identification. In *ISMIR*, pages 149–154, 2013.

Hoon Heo, Hyunwoo J Kim, Wan Soo Kim, and Kyogu Lee. Cover song identification with metric learning using distance as a feature. In *ISMIR*, 2017.

Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM, 2007.

Ning Hu, Roger B Dannenberg, and George Tzanetakis. Polyphonic audio matching and alignment for music retrieval. In *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pages 185–188. IEEE, 2003.

Joan Serra, Emilia Gómez, and Perfecto Herrera. Transposing chroma representations to a common key. In *IEEE CS Conference on The Use of Symbols to Represent Music and Multimedia Objects*, pages 45–48, 2008b.

Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.

Jouni Paulus, Meinard Müller, and Anssi Klapuri. State of the art report: Audio-based music structure analysis. In *ISMIR*, pages 625–636, 2010.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.