

# LEARNING A REPRESENTATION FOR COVER SONG IDENTIFICATION USING CONVOLUTIONAL NEURAL NETWORK

Zhesong Yu, Xiaoshuo Xu, Xiaoou Chen, Deshun Yang

Wangxuan Institute of Computer Technology, Peking University

## ABSTRACT

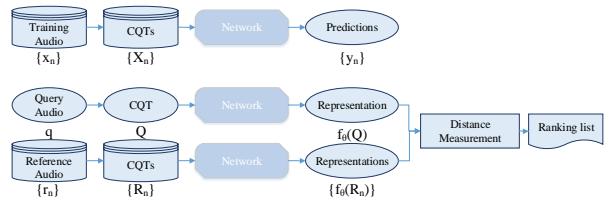
Cover song identification represents a challenging task in the field of Music Information Retrieval (MIR) due to complex musical variations between query tracks and cover versions. Previous works typically utilize hand-crafted features and alignment algorithms for the task. More recently, further breakthroughs are achieved employing neural network approaches. In this paper, we propose a novel Convolutional Neural Network (CNN) architecture based on the characteristics of the cover song task. We first train the network through classification strategies; the network is then used to extract music representation for cover song identification. A scheme is designed to train robust models against tempo changes. Experimental results show that our approach outperforms state-of-the-art methods on all public datasets, improving the performance especially on the large dataset.

**Index Terms**— Music Information Retrieval, Cover Song Identification

## 1. INTRODUCTION

Cover song identification has long been a popular task in the music information retrieval community, with potential applications in areas such as music license management, music retrieval, and music recommendation. Cover song identification can also be seen as measuring the similarity between music melodies. Given those cover songs may differ from the original song in key transposition, speed change and structural variations, identifying cover songs is a rather challenging task. Over the past ten years, researchers initially attempt to address the problem employing Dynamic Programming (DP) approaches. Typically, chroma sequences representing the intensity of twelve pitch classes are used to describe recordings, and then a DP method is utilized for finding an optimal alignment between two given recordings, resolving the discrepancy caused by tempo changes and structural variations [1, 2, 3]. Such approaches work well when facing structural variations and tempo changes in the music; however, the involvement of element-to-element distance computation with quadratic time complexity makes it unsuitable for large-scale datasets.

Alternatively, some researchers attempted to identify



**Fig. 1.** Training procedure and retrieval procedure.

cover songs by modeling the music. For instance, Serrà et al. studied time series modeling for cover song identification [3]. [4, 5] represent the music with fixed-dimensional vectors, which enables a direct measure of the music similarity. These approaches highly improved the efficiency compared to alignment methods, while the loss of the temporal information of music in these approaches may yield a lower precision.

Moreover, deep learning approaches are introduced to cover song identification. For instance, CNNs are utilized to measure the similarity matrix [6] or learn features [7, 8, 9, 10]. While these methods have achieved promising results, there is still room for improvement. In this paper, a specially designed CNN architecture is proposed to overcome challenges of key transposition, speed change and structural variations existing in cover song identification. Notably, the use of the specialized kernel size is first ever utilized in the field of music information retrieval. The dilation convolution and the method of data augmentation are also introduced. Our approach outperforms state-of-the-art methods on all public datasets with better accuracy but lower time complexity. Furthermore, to our best knowledge, our method is currently the best method to identify cover songs in huge real-life corpora.

## 2. APPROACH

### 2.1. Problem Formulation

As shown in Figure 1, we have a training dataset  $D = \{(x_n, t_n)\}$ , where  $x_n$  is a recording and  $t_n$  is a one-hot vector denoting to which song (or class) the recording belongs. Different versions of the same song are viewed as the samples from the same class, and different songs are regarded as the different classes. We aim to train a classification net-

work model parameterized as  $\{\theta, \lambda\}$  from  $D$ . As shown in Figure 2,  $\theta$  is the parameter of all convolutions and FC0;  $\lambda$  is the parameter of FC1;  $f_\theta$  is the output of the FC0 layer. Then, this model could be used for cover song retrieval. More specifically, after the training, given a query  $Q$  and references  $R_n$  in the dataset, we extract latent features  $f_\theta(Q), f_\theta(R_n)$  which we call as music representations obtained by the network, and then we use a metric  $s$  to measure the similarity between them. In the following sections, we will discuss the low-level representation used, the design of network structure and a robust trained model against key transposition and tempo change. We use lowercase for audio and uppercase for the CQT to discriminate.

## 2.2. Low-level Representation

The CQT, mapping frequency energy into musical notes, is extracted by *Librosa* [11] for our experiment. The audio is resampled to 22050 Hz, the number of bins per octave is set as 12 and Hann window is used for extraction with a hop size of 512. Finally, a 20-point mean down-sampling in the time direction is applied to the CQT, and the resulting feature is a sequence with a feature rate of about 2 Hz. It could also be viewed as an  $84 \times T$  matrix where  $T$  depends on the duration of input audio.

## 2.3. Network Structure

Inspired by successful applications of network architectures like [12, 13], we design a novel Network architecture for the cover song task. We stack small filters following with max pooling operations, except that in initial layers, we design the height of filter to be 12 or 13 (see Figure 2) as the number of bins per octave is set as 12 in the CQT. This setting results in that the units of the third layer have a receptive field with a height of 36; it spans across three octaves or thirty-six semitones.

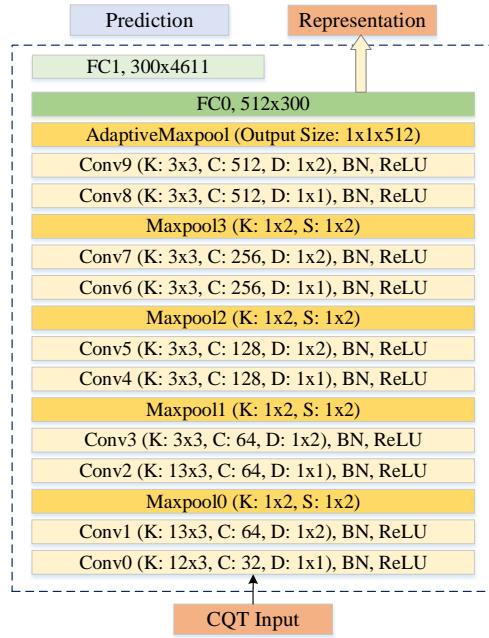
We also introduce dilated convolution into the model to enlarge the receptive field because cover song identification focuses on the long-term melody of the music. The design is consistent with the ideas of existing works in [2, 4], which extracted features or measured the similarity from a long range.

More importantly, our model does not involve any down-sample pooling operation in the frequency dimension; in other words, the vertical stride is always set to 1, different from prevalent network structures like VGG and ResNet [12, 13]. The motivation behind this design focuses on the fact that key transposition may be one or two semitones, corresponding to moving the CQT matrix vertically for merely one or two elements. Without downsampling feature map, the network remains a higher resolution and deals with key transposition better. We experimentally validate that this design helps improve the precision (see Section 4.1).

Furthermore, after several convolutional and pooling layers, we apply an adaptive pooling layer to the feature map,

whose length varies depending on the input audio. Obviously, this global pooling has the advantage of converting variable-length feature maps into fixed-dimensional vectors, connected with two fully-connected layers to exploit more information. Without the global pooling, the fully-connected layers only allow fixed-dimensional inputs, which are not the cases of music as different compositions may have different durations. Comparing with [9] who utilized Temporal Pyramid Pooling, we utilize global pooling because it performs the same as TPP in our model. One explanation is that our convolutional network is much deeper.

Given the inputs of network  $X$ , the output of FC0 is  $f_\theta(X) \in R^{300}$ , and the prediction of network is  $y = \text{softmax}(\lambda f_\theta(X))$ . Cross-entropy loss  $L$  is used for training. In our cases, different versions of a song are considered as the same category, and different songs are viewed as different classes.



**Fig. 2.** Network structure. K: kernel size, C: channel number, D: dilation and S: stride. The stride is set to  $1 \times 1$  for the convolutional layers, and pooling layers has a dilation of  $1 \times 1$ . The output dimension is 4611, the number of classes in the training set.

## 2.4. Training Scheme

For each batch, we sample some recordings from the training set and extract CQTs from them. For each CQT, we randomly crop three subsequences with a length of 200, 300 and 400 for training, corresponding to roughly 100s, 150s and 200s, respectively.

---

**Algorithm 1** Data augmentation and training strategy

---

**Input:** Training set  $D$ , batch size  $n$ , changing range  $(a, b)$   
**Output:** Optimized parameter  $\theta, W$

```
1: repeat
2:   for  $L \in \{200, 300, 400\}$  do
3:     sample a batch of recordings  $B$  from the training set
4:      $D$ 
5:      $S \leftarrow \emptyset$ 
6:     for  $x \in B$  do
7:        $r \leftarrow$  sample from  $U(a, b)$ 
8:        $x \leftarrow$  simulate tempo changes on  $x$  with a chang-
ing factor  $r$ 
9:        $X \leftarrow$  extract the CQT from  $x$ 
10:       $X \leftarrow$  crop a subsequence from  $X$  with a length  $l$ 
11:       $S \leftarrow S \cup X$ 
12:    end for
13:    Feed-forward with  $T$ 
14:    Backpropagation to update  $\theta$  and  $W$ 
15:  end for
15: until Network converges
```

---

Despite the training set contains covers performed at different speeds, each song merely owns several covers on average for training. Moreover, as our model does not explicitly handle tempo changes in cover songs, it may be difficult for the model to learn a representation robust against tempo changes automatically. Therefore, we perform data augmentation during model training. As shown in Algorithm 1, we sample a changing factor from  $(0.7, 1.3)$  for each recording in the batch following uniform distribution and simulate tempo changes using *Librosa* [11] on the recording before cropping subsequences.

## 2.5. Retrieval

After the training, the network is used to extract music representations. As shown in Figure 1, given a query  $q$  and a reference  $r$ , we first extract their CQT descriptors  $Q$  and  $R$  respectively, which are fed into the network to obtain music representations  $f_\theta(Q)$  and  $f_\theta(R)$ , and then the similarity  $s$ , defined as their cosine similarity, are measured. After computing the pair-wise similarity between the query and references in the dataset, a ranking list is returned for evaluation.

## 3. EXPERIMENTAL SETTINGS

### 3.1. Dataset

*Second Hand Songs 100K (SHS100K)*, which is collected from *Second Hand Songs website* by [8], consisting of 8858 songs with various covers and 108523 recordings. This dataset is split into three subsets – *SHS100K-TRAIN*, *SHS100K-VAL* and *SHS100K-TEST* with a ratio of 8 : 1 : 1.

*Youtube* is collected from the YouTube website, containing 50 compositions of multiple genres [14]. Each song in *Youtube* has 7 versions, with 2 original versions and 5 different versions and thus results in 350 recordings in total. In our experiment, we use the 100 original versions as references and the others as queries following the same as [15, 9, 8].

*Covers80* is a widely used benchmark dataset in the literature. It has 80 songs, with 2 covers for each song, and has 160 recordings in total. To compare with existing methods, we compute the similarity of any pair of recordings.

*Mazurkas* is a classical music collection consisting of 2914 recordings of 49 Chopin’s Mazurkas, originated from the Mazurka Project<sup>1</sup>. The number of covers for each piece varies between 41 and 95. For this dataset, we follow the experimental setting of [15].

### 3.2. Evaluation

For evaluation, we calculate the common evaluation metrics: mean average precision (MAP), precision at 10 (P@10) and the mean rank of the first correctly identified cover (MR1). These metrics are the ones used in Mirex Audio Cover Song Identification contest<sup>2</sup>. Additionally, query time is recorded for efficiency evaluation. All the experiments are run in a Linux server with two TITAN X (Pascal) GPUs.

## 4. EXPERIMENTAL RESULT AND ANALYSIS

### 4.1. Exploration of Network Structure

Firstly we explore the kernel size of CNNs and replace the kernel size of the initial three layers with  $3 \times 3$ ,  $7 \times 3$ ,  $15 \times 3$ ,  $7 \times 7$ ,  $12 \times 12$ . The result of experiment shows that the height of filter to be 12 or 13 performs the best.

Additional, we change the vertical strides of max-pooling layers and conduct several experiments to explore its influence on accuracy. The original model is denoted as CQT-Net, and the modified network is denoted as CQT-Net{4}, CQT-Net{3, 4} and CQT-Net{2, 3, 4} respectively, where the numbers indicate the shape of corresponding pooling layers are changed to  $(2, 2)$ . For instance, CQT-Net{2} means that Pool2 is replaced with a pooling operation with a stride and size of  $(2, 2)$ . That is, the total vertical strides for CQT-Net{4}, CQT-Net{3, 4} and CQT-Net{2, 3, 4} are 2, 4 and 8, respectively. When the vertical stride increases, MAP degrades on the four datasets consistently, as well as MR1 and P@10. We suppose this is because the key transposition may shift one or two semitones; the network having a higher resolution of feature dimension (that is, setting vertical stride to be 1) could capture these changes and help improve the precision.

---

<sup>1</sup>[www.mazurka.org.uk](http://www.mazurka.org.uk)

<sup>2</sup>[https://www.music-ir.org/mirex/wiki/2019:Audio\\_Cover\\_Song\\_Identification](https://www.music-ir.org/mirex/wiki/2019:Audio_Cover_Song_Identification)

	MAP	P@10	MR1	Time
<i>Results on Youtube</i>				
DPLA [2]	0.525	0.132	9.43	2420s
SiMPle [15]	0.591	0.140	7.91	18.7s
Fingerprinting [16]	0.648	0.145	8.27	-
SuCo-DTW [17]	0.800	0.180	3.42	4.59s
Ki-CNN [8]	0.656	0.155	6.26	0.35ms
TPPNet [9]	0.859	0.188	2.85	0.04ms
CQT-Net	<b>0.917</b>	<b>0.192</b>	<b>2.50</b>	<b>0.04ms</b>
<i>Results on Covers80</i>				
NCP-WIDI [18]	0.645	-	-	-
CRP [3]	0.544	0.061	-	-
Fusing [19]	0.625	0.071	-	-
Ki-CNN [8]	0.506	0.068	16.4	0.55ms
TPPNet [9]	0.744	0.086	6.88	0.06ms
CQT-Net	<b>0.840</b>	<b>0.091</b>	<b>3.85</b>	<b>0.06ms</b>
<i>Results on Mazurkas</i>				
DTW [15]	0.882	0.949	4.05	-
NCD [20]	0.767	-	-	-
Compression [21]	0.795	-	-	-
Fingerprinting [22]	0.819	-	-	-
SiMPle [15]	0.880	0.952	2.33	-
SuCo-repeat [17]	0.850	0.940	2.77	-
2DFM [4]	0.363	0.578	15.6	4.76ms
Ki-CNN [8]	0.707	0.892	4.01	5.34ms
CQT-Net	<b>0.933</b>	<b>0.956</b>	<b>2.87</b>	<b>0.50ms</b>
<i>Results on SHS100K-TEST</i>				
2DFM [4]	0.104	0.113	415	13.9ms
Ki-CNN [8]	0.219	0.204	174	21.0ms
TPPNet [9]	0.465	0.357	72.2	3.68ms
CQT-Net	<b>0.655</b>	<b>0.456</b>	<b>54.9</b>	<b>3.68ms</b>

**Table 1.** Performance on different datasets (- indicates the results are not shown in original works).

## 4.2. Comparison

We compare with other state-of-the-art methods on different datasets. Table 4.1 shows our approach outperforms state-of-the-art methods on all datasets. The advantages of our approach lie in without relying on complicated hand-crafted features and elaborately-designed alignment algorithms, our approach exploits massive data and feature learning and obtains high precision. By collecting a larger dataset, our approach may obtain higher precision. It is worth noting that our training set *SHS100K-TRAIN* mainly consists of pop music while *Mazurkas* contains classical music. Our approach outperforms state-of-the-art methods on this dataset, which indicates a good generalization ability. We do not show the result of [22] in *Mazurka* Project because our test sets are different. As for the large dataset *SHS100K-TEST*, our method performs much better than state-of-the-art methods.

Moreover, the query time shown in the table does not include the time of feature extracting. Therefore, our method has the same time consumed as [9]. It extracts a fixed-dimensional feature whatever the duration of input audio

is. Theoretically, it has linear time complexity, faster than sequence alignment methods with quadratic time complexity. One could find that the query time of our approach is shorter than approaches such as DPLA, SiMPle by several magnitudes. For Ki-CNN and TPPNet, they model music with a fixed-dimensional vector and have similar time complexity. In our implementation, our approach learns a 300-dimensional representation, which is the same as TPPNet, explaining why the time consumption of our approach is the same as TPPNet.

## 4.3. Result Demonstration and Error Analysis

We listen to the Top10 retrieval results and attempt to make some analysis on *SHS100K-TEST*. Our approach could identify versions when performed by different genders, accompanied by different instruments, sung in different languages, etc. Especially, as our training goal is to classify the song and different versions of the same song often have similar styles, melodic and chord structures, we find that even though some candidates in the Top10 may not be the cover of the query, but they have similar properties such as accompaniment and genre with the query. For instance, *Everybody Knows This Is Nowhere* by the Bluebeaters has a similar accompaniment with that of *Waiting in Vain* by Bob Marley & The Wailers. In this sense, our approach may also be used to retrieve similar music of the query and extended to content-based music recommendation.

Furthermore, we find that Top1 precision of our model is 0.81, suggesting that it could find a cover as the Top1 candidate for 81% queries. However, it works worse in some cases. This may explain why our approach obtains a MAP of 0.655 while only a MR1 of 54.9 on this dataset. Most importantly, the high Top1 precision and the fast retrieval speed make our method possible to handle the real-life cover song task instead of just staying in the lab stage.

## 5. CONCLUSION

Different from conventional techniques, we propose CNNs for feature learning towards cover song identification. Utilizing specific kernels and dilated convolutions to extend the receptive field, we show that it could be used to capture melodic structures underlying the music and learn key-invariant representations. By casting the problem into a classification task, we train a model that is used for music version identification. Additionally, we design a training strategy to enhance the model’s robustness against tempo changes and to deal with inputs with different lengths. Combined with these techniques, our approach outperforms state-of-the-art methods on all public datasets with low time complexity. Furthermore, we show that this model could retrieve various music versions and discover similar music. Eventually, we believe our method is competent to solve real-life cover song problem.

## 6. REFERENCES

- [1] Daniel PW Ellis and Graham E Poliner, “Identifying cover songs with chroma features and dynamic programming beat tracking,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.
- [2] Joan SerrÃ , Emilia Gómez, Perfecto Herrera, and Xavier SerrÃ , “Chroma binary similarity and local alignment applied to cover song identification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1138–1151, 2008.
- [3] Joan SerrÃ , Xavier SerrÃ , and Ralph G Andrzejak, “Cross recurrence quantification for cover song identification,” *New Journal of Physics*, vol. 11, no. 9, pp. 093017, 2009.
- [4] Thierry Bertin-Mahieux and Daniel PW Ellis, “Large-scale cover song recognition using the 2d fourier transform magnitude,” in *International Society for Music Information Retrieval Conference*, 2012.
- [5] Julien Osmalsky, Marc Van Droogenbroeck, and Jean-Jacques Embrechts, “Enhancing cover song identification with hierarchical rank aggregation,” in *International Society for Music Information Retrieval Conference*, 2016, pp. 136–142.
- [6] Sungkyun Chang, Juheon Lee, Sang Keun Choe, and Kyogu Lee, “Audio cover song identification using convolutional neural network,” in *Workshop Machine Learning for Audio Signal Processing at NIPS*, 2017.
- [7] Xiaoyu Qi, Deshun Yang, and Xiaou Chen, “Audio feature learning with triplet-based embedding network.,” in *AAAI*, 2017, pp. 4979–4980.
- [8] Xiaoshuo Xu, Xiaou Chen, and Deshun Yang, “Key-invariant convolutional neural network toward efficient cover song identification,” in *2018 IEEE International Conference on Multimedia and Expo*. IEEE, 2018, pp. 1–6.
- [9] Zhesong Yu, Xiaoshuo Xu, Xiaou Chen, and Deshun Yang, “Temporal pyramid pooling convolutional neural network for cover song identification,” in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 4846–4852.
- [10] Guillaume Doras and Geoffroy Peeters, “Cover detection using dominant melody embeddings,” *CoRR*, vol. abs/1907.01824, 2019.
- [11] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, “librosa: Audio and music signal analysis in python,” in *Proceedings of the 14th python in science conference*, 2015, pp. 18–25.
- [12] Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *International Conference on Learning Representations*, 2015.
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [14] Diego Furtado Silva, Vinícius MA de Souza, and Gustavo EAPA Batista, “Music shapelets for fast cover song recognition.,” in *International Society for Music Information Retrieval Conference*, 2015, pp. 441–447.
- [15] 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*, 2016.
- [16] Prem Seetharaman and Zafar Rafii, “Cover song identification with 2d fourier transform sequences,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 616–620.
- [17] Diego Furtado Silva, Felipe Vieira Falcao, and Nazareno Andrade, “Summarizing and comparing music data and its application on cover song identification,” in *International Society for Music Information Retrieval Conference*, 2018.
- [18] Yao Cheng, Xiaou Chen, Deshun Yang, and Xiaoshuo Xu, “Effective music feature ncp: Enhancing cover song recognition with music transcription,” in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2017, pp. 925–928.
- [19] Ning Chen, Wei Li, and Haidong Xiao, “Fusing similarity functions for cover song identification,” *Multimedia Tools and Applications*, vol. 77, no. 2, pp. 2629–2652, 2018.
- [20] Juan P Bello, “Measuring structural similarity in music,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2013–2025, 2011.
- [21] Diego Silva, Hélène Papadopoulos, Gustavo EAPA Batista, and Daniel PW Ellis, “A video compression-based approach to measure music structural similarity,” in *14th International Society for Music Information Retrieval Conference*, 2013, pp. 95–100.
- [22] Peter Grosche, Joan SerrÃ , Meinard Müller, and Josep Lluis Arcos, “Structure-based audio fingerprinting for music retrieval.,” in *International Society for Music Information Retrieval Conference*, 2012, pp. 55–60.