Improved Similarity Fusion Scheme for Cover Song Identification.

Y.L. Fan and N. Chen

A Cover Song Identification (CSI) scheme based on non-linear graph fusion and Tensor Product Graphs (TPGs) diffusion is proposed as an improvement to our previously proposed similarity fusion-based CSI scheme. First, the harmonic progression, melody evolution, and rhythm based descriptors are extracted from the track, respectively. Next, Similarity Network Fusion (SNF) is adopted to fuse the similarity graphs obtained based on two types of descriptors to take full use of the common as well as complementary properties between them. Finally, TPGs diffusion is performed on the obtained fused similarity graphs to take advantage of the manifold structure contained in them to improve the performance, further. Experimental results demonstrate the superiority of the proposed scheme over our previously proposed one, in terms of identification accuracy and clustering performance.

Introduction: The Cover Song Identification (CSI) technique is to identify the different versions, performances, or renditions of a specific track. It is a challenging task because the cover version may differ from the original one in various musical aspects, such as timbre, tempo, key, structure, lyrics and language, and so on. Since CSI technique can be applied in music collection organization, music rights management and licensing, and music creation aiding, it has become an active studying area in music information retrieval.

To enhance the CSI performance, several similarity fusion algorithms [1, 2] were put forward to take advantage of the common as well as complementary properties among different descriptors (such as the Harmonic Pitch Class Profile (HPCP) [3], MeLoDy (MLD) [4], Beat-Synchronous Chroma (BSC) [5]). For example, in [1], a non-linear graph fusion technique was adopted to fuse the similarity graphs constructed based on different descriptors (HPCP, Cochlear Pitch Class Profile (CPCP) and BSC). However, since some important factors that may influence the performances greatly were not considered in [1], its performances were affected. First, since both CPCP and HPCP describe the harmonic progression property of the track, the complementarity between them is limited. So, the fusion of CPCP and HPCP has little contribution to enhancing the performance. Second, the scheme in [1] fuses the similarities based on three descriptors, directly, which does not take full advantage of the complementarity between any two descriptors. Third and most important, the track manifold structure contained in the fused similarity graph, which can be used to reduce the influence of noise, is ignored in [1].

To solve these possible problems and enhance the performance, further, an improved version of the scheme in [1] is proposed in this letter. First, the MLD descriptor, which describes the main melody evolution property of the track, is adopted to replace CPCP descriptor. Thus, the HPCP, which describes the harmonic progression, the MLD, which describes the melody evolution, and the BSC, which represents the rhythm property, have high complementarity between each other. Second, the Similarity Network Fusion (SNF) [6] is used to fuse the similarity graphs constructed based on any two descriptors, which helps to fully utilize the common as well as complementary properties between them. Third, the Tensor Product Graphs (TPGs) diffusion technique [7] is adopted to take advantage of the track manifold structures contained in each SNF fused similarity graph to reduce the noise influence. Extensive experiments demonstrate the superiority of the proposed scheme over the scheme in [1] and other CSI schemes based on single similarity or similarity fusion, in terms of cover song identification accuracy and cover song dataset clustering accuracy.

Proposed scheme: The block diagram of the proposed similarity fusion scheme is shown in Fig. 1.

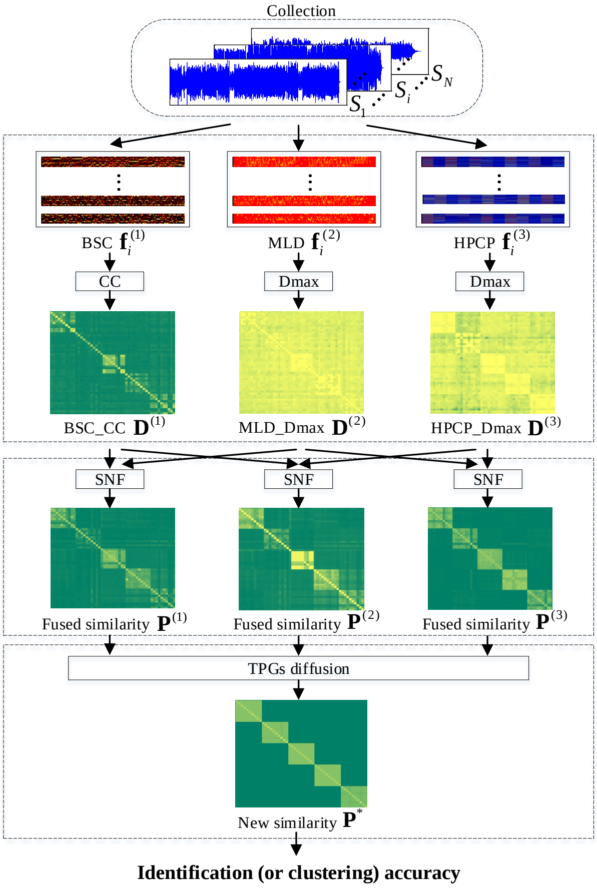
Collection
S_1
S_i
S_N
BSC $\mathbf{f}_i^{(1)}$
MLD $\mathbf{f}_i^{(2)}$
HPCP $\mathbf{f}_i^{(3)}$
CC
Dmax
Dmax
BSC_CC $\mathbf{D}^{(1)}$
MLD_Dmax $\mathbf{D}^{(2)}$
HPCP_Dmax $\mathbf{D}^{(3)}$
SNF
SNF
SNF
Fused similarity $\mathbf{P}^{(1)}$
Fused similarity $\mathbf{P}^{(2)}$
Fused similarity $\mathbf{P}^{(3)}$
TPGs diffusion
New similarity $\mathbf{P}^*$
Identification (or clustering) accuracy

Fig. 1. Block diagram of the proposed CSI scheme.

- **Feature extraction and similarity calculation:** Assume the track collection, denoted as $\mathbf{S} = \{S_1, \cdots, S_i, \cdots, S_N\}$, is composed of N tracks. For each track S_i, three descriptors: BSC, denoted as $\mathbf{f}_i^{(1)}$, MLD, denoted as $\mathbf{f}_i^{(2)}$, and HPCP, denoted as $\mathbf{f}_i^{(3)}$, are extracted, respectively.

  In the proposed scheme, the Dmax [8] and Cross-Correlation (CC) [5] are adopted to meassure the similarity between HPCP (or MLD) descriptors and BSC descriptors, respectively. For $\mathbf{S}$, the similarity matrices based on BSC, MLD, and HPCP descriptors are denoted as $\mathbf{D}^{(1)}$, $\mathbf{D}^{(2)}$ and $\mathbf{D}^{(3)}$, respectively.
- **Cross similarity matrices fusion:** To take full advantage of the common as well as complementary properties between any two types of descriptors, SNF technique [6] is adopted to fuse the combination of any two similarity matrices ($\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$, $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(3)}$, $\mathbf{D}^{(2)}$ and $\mathbf{D}^{(3)}$) to obtain the cross-fused similarity matrices, denoted as $\mathbf{P}^{(1)}$, $\mathbf{P}^{(2)}$, and $\mathbf{P}^{(3)}$, respectively.
- **Diffusion based on TPGs:** To take full advantage of the track manifold structures contained in the cross-fused similarity matrices $\{\mathbf{P}^{(i)} \in \mathbb{R}^{N\times N}, i = 1, 2, 3\}$, TPGs-based diffusion is performed on them as follows.

  First, the Kronecker product, denoted as $\bigotimes$, is applied on any two different cross-fused similarity matrices $\mathbf{P}^{(i)}$ and $\mathbf{P}^{(j)}$ to obtain $\mathbb{P}_{(i,j)} \in \mathbb{R}^{NN\times NN}$ with Eq. (1).

$$\mathbb{P}_{(i,j)} = \mathbf{P}^{(i)} \bigotimes \mathbf{P}^{(j)}, \quad i, j = 1, 2, 3 \quad \& \quad i \neq j \qquad (1)$$

  Next, all $\mathbb{P}_{(i,j)}$, are added to obtain the TPGs $\mathbb{P}$ with Eq. (2).

$$\mathbb{P} = \sum_{i=1}^{3}\sum_{j=1}^{3} \mathbb{P}_{(i,j)}, \quad i, j = 1, 2, 3 \quad \& \quad i \neq j \qquad (2)$$

  Then, the diffusion on $\mathbb{P}$ can be defined with Eq. (3).

$$\mathbb{P}^{(t)} = \sum_{k=1}^{t} \mathbb{P}^k \qquad (3)$$

where t is the iteration time. The nontrivial solution, denoted as $\mathbb{P}^*$, of Eq. (3) can be obtained with Eq. (4).

$$\mathbb{P}^* = \lim_{t\to\infty} \mathbb{P}^{(t)} = \lim_{t\to\infty} \sum_{k=1}^{t} \mathbb{P}^k = (\mathbf{I} - \mathbb{P})^{-1} \qquad (4)$$

where $\mathbf{I}$ is an identity matrix.

Finally, the refined similarity matrix, denoted as $\mathbf{P}^* \in \mathbb{R}^{N\times N}$, can be obtained from $\mathbb{P}^*$ with Eq. (5).

$$\mathbf{P}^* = vec^{-1}(\mathbb{P}^* vec(\mathbf{I})) = vec^{-1}((\mathbf{I} - \mathbb{P})^{-1} vec(\mathbf{I})) \quad (5)$$

where vec is an operator stacking columns of a matrix one by one into a column vector.

In addition, considering that the TPGs diffusion process shown in Eq. (3) requires much storage and computation cost, it can be optimized as Eq. (6).

$$Q^{(t+1)} = (\sum_{i=1}^{3} \mathbf{P}^{(i)}) Q^t (\sum_{j=1}^{3} \mathbf{P}^{(j)})^T + \mathbf{I} \quad (6)$$

where $Q^{(1)} = \sum_{i=1}^{3} \mathbf{P}^{(i)}$. Then the $\mathbf{P}^*$ can be calculated with Eq. (7).

$$\mathbf{P}^* = \lim_{t\to\infty} Q^{(t)} \quad (7)$$

$\mathbf{P}^*$ is the learned new similarity, which can be used to CSI task.

Experimental results: To verify the superiority of the proposed scheme (called as CSNF-TPGs) over state-of-the-art similarity fusion based CSI ones, the performances of the proposed scheme, in terms of identification accuracy and clustering accuracy, are compared with those of schemes in [1] and [2] on three datasets (Covers80[1], Covers40, and Covers4235). Covers80 contains 80 groups of cover songs and each group has two songs. Covers40 contains 40 groups of cover songs and 10 songs per group. Covers4235, which is a part of Second Hand Song (SHS)[2] dataset, is composed of 12730 tracks which are classified into 4235 groups. Three evaluation measures, the Mean of Average Precision (MAP) [9], the total number of identified covers in the Top 10 (Top-10), and the Mean averaged Reciprocal Rank (MaRR) [10], are adopted to evaluate the identification accuracy. It should be noted that for the schemes in [1] and [2], the similarities based on the same descriptors as those adopted by the proposed scheme are fused.

The identification accuracy comparison results shown in Table 1 demonstrate that: i) For the proposed scheme, the cross-fused similarity performs better than the fusion objects in terms of all evolution methods on all three datasets. ii) The TPGs diffusion helps to enhance the performance further. iii) The proposed scheme outperforms the similarity fusion-based schemes in [1] and [2] in terms of all evolution methods on all three datasets.

Table 1: Identification accuracy comparison results.

Datasets	Schemes	MAP	TOP10	MaRR
Covers80	BSC_CC [5]	0.4506	80	0.2394
	MLD_Dmax	0.3283	67	0.1818
	HPCP_Dmax	0.5709	104	0.2979
	SNF (BSC_CC+MLD_Dmax)	0.6053	108	0.3125
	SNF (MLD_Dmax+HPCP_Dmax)	0.6672	122	0.3462
	SNF (BSC_CC+HPCP_Dmax)	0.7062	125	0.3646
	[2]	0.6041	108	0.3156
	[1]	0.7451	129	0.3815
	CSNF-TPGs	**0.7547**	**130**	**0.3843**
Covers40	BSC_CC [5]	0.4363	1523	0.1231
	MLD_Dmax	0.4867	1654	0.1362
	HPCP_Dmax	0.7945	2717	0.1908
	SNF (BSC_CC+MLD_Dmax)	0.8081	2710	0.1874
	SNF (MLD_Dmax+HPCP_Dmax)	0.9710	3456	0.2098
	SNF (BSC_CC+HPCP_Dmax)	0.8415	2829	0.1911
	[2]	0.7770	2625	0.1867
	[1]	0.9368	3314	0.2043
	CSNF-TPGs	**0.9850**	**3528**	**0.2117**
Covers4235	BSC_CC [5]	0.0513	4743	0.0353
	MLD_Dmax	0.1612	5790	0.0794
	HPCP_Dmax	0.3927	13974	0.1821
	SNF (BSC_CC+MLD_Dmax)	0.2052	5984	0.0977
	SNF (MLD_Dmax+HPCP_Dmax)	0.4519	16014	0.2015
	SNF (BSC_CC+HPCP_Dmax)	0.4479	16034	0.2020
	[2]	0.2367	9611	0.1199
	[1]	0.4482	15901	0.2025
	CSNF-TPGs	**0.4648**	**17694**	**0.2032**

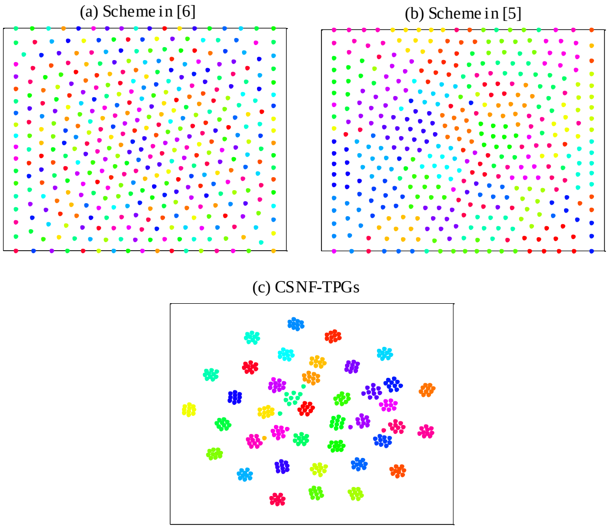

Fig. 2. Clustering performance comparison results on Covers40.

As shown in Fig. 2, the clustering performances of three schemes (CSNF-TPGs, and those in [1] and [2]) are only compared on Covers40, but not Covers80 or Covers4235. The reason is that the sizes of the cover sets in Covers80 and Covers4235 are too large to be visualized clearly. It can be seen that the proposed scheme performs better than the schemes in [1] and [2] because the intra-distances are much smaller than the inter-distances in the proposed scheme.

Conclusions: A modified version of the scheme in [1] is proposed for CSI task. In the proposed scheme, the complementarity between any two descriptors are taken full advantage of. In addition, diffusion process based on TPGs is performed on the cross-fused similarity matrices to enhance the performance, further. Experimental results demonstrate that the proposed scheme outperforms state-of-the-art CSI schemes based on single similarity or similarity fusion.

Acknowledgment: This work was supported by the National Natural Science Foundation of China (61771196).

Y.L. Fan (*School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China.*)

N. Chen (*School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China.*)

E-mail: chenning_750210@163.com

References

1 Chen N, Xiao H.:'Similarity fusion scheme for cover song identification'. *Electronics Letters*, 2016, **52(13)**, pp. 1173-1175.

2 Degani A, Dalai M, Leonardi R, et al.: 'A heuristic for distance fusion in cover song identification', Image Analysis for Multimedia Interactive Services (WIAMIS), 2013 14th International Workshop on. IEEE, 2013, pp. 1-4.

3 Gomez, E.:'Tonal description of music audio signals', *Phd Dissertation of Universitat Pompeu Fabra*, 2006.

4 Serra, J.: 'Identification of versions of the same musical composition by processing audio descriptions', *Department of Information and Communication Technologies*, 2011.

5 Ellis, D. P.: 'Identifying 'cover songs' with beat-synchronous chroma features', *MIREX*, 2006, pp. 1-4.

6 Wang, B., Mezlini, A.M., Demir, F., Fiume, M. Tu, Z.W., Brudno, M. Haibe-Kains, B. and Goldenberg, A.:'Similarity network fusion for aggregating data types on a genomic scale', *Nature methods*, 2014, **11:3**, pp. 333-337.

7 Shu L, Latecki L J.: 'Integration of single-view graphs with diffusion of tensor product graphs for multi-view spectral clustering', *Asian Conference on Machine Learning*. 2016, pp. 362-377.

8 Yang, F., Chen, N.: 'Cover song identification based on cross recurrence plot and local alignment', *Journal of East China University of Science and Technology*, 2016, **42(2)**, pp. 247-253.

9 Serra, J., Serra, X. and Andrzejak, R. G.: 'Cross recurrence quantification for cover song identification', *New J. Physics*, 2009, **3:9**, pp. 093017.

10 Salamon, Justin J.: 'Melody extraction from polyphonic music signals', 2013.

[1] https://labrosa.ee.columbia.edu/projects/coversongs/covers80/

[2] https://secondhandsongs.com/