Chen *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2017) 2017:12
DOI 10.1186/s13636-017-0108-2

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Two-layer similarity fusion model for cover song identification

Ning Chen[1*], Mingyu Li[1] and Haidong Xiao[2]

Abstract

Various musical descriptors have been developed for Cover Song Identification (CSI). However, different descriptors are based on various assumptions, designed for representing distinct characteristics of music, and often differ in scale and noise level. Therefore, a single similarity function combined with a specific descriptor is generally not able to describe the similarity between songs comprehensively and reliably. In this paper, we propose a two-layer similarity fusion model for CSI, which combines the information carried by different descriptors and similarity functions organically and incorporates the advantages of both early fusion and late fusion. In particular, in the early fusion, the similarities obtained by the same descriptor and different similarity functions are integrated with the Similarity Network Fusion (SNF) technique. Then, in the late fusion, the learning method selected by sparse group LASSO algorithm is applied on each early fused similarity to obtain the probability that the corresponding song pair belongs to the reference/cover pair. Lastly, the final fused similarity is achieved by averaging all the obtained probabilities. Extensive experimental results on the music collection that is composed of samples provided by the SecondHandSongs (SHS) verify that the proposed scheme outperforms state-of-the-art fusion based CSI schemes in terms of identification accuracy and classification efficiency.

Keywords: Cover Song Identification (CSI), Music Information Retrieval (MIR), Early fusion, Late fusion, Two-layer similarity fusion

1 Introduction

The explosion of musical data makes us face new challenges unthinkable two decades ago. For example, how to retrieve different versions, performance, or renditions of a previously recorded musical composition has become a challenging problem [1]. Cover Song Identification (CSI) can help in this regard. Its potential applications include music right and licenses management and music creation aid. It has become an active research field in Music Information Retrieval (MIR) over the past decades.

Since the cover version may be obtained in different ways (such as remastering, instrumental, mashup, live performance, acoustic, demo, remix, quotation, medley, and standard [2]), it may differ from the original in timbre, tempo, timing, structure, key, harmonization, lyrics and language, and/or noise [3]. What remains almost invariable among the various cover versions are harmonic progressions and melody evolution, which form the basis of the most existing CSI descriptor extraction algorithms. Among these descriptors, the Chroma (also called Pitch Class Profiles (PCP)) [4] and its variations [5–13] are the most widely-used descriptors for describing harmonic progressions. In [9], the beat-synchronous chroma for two tracks were cross-correlated, from the results of which the sharp peaks indicating good local alignment were looked for to determine the distance between them. It performed the best in the audio CSI task contest of the 2006 Music Information Retrieval Evaluation eXchange (MIREX) [14]. The Harmonic PCP (HPCP) descriptor proposed in [15] shares the common properties of PCP, but since it is only based on the peaks of the spectrum within a certain frequency band, it reduces the influence of noisy spectral components further. It also takes the presence of harmonic frequencies into account and is tuning independent. The CSI scheme based on HPCP and $Q_{\max}$ similarity measure [5, 16] achieved the highest identification accuracy in the 2009 MIREX audio CSI task contest. In [10],

*Correspondence: chenning_750210@163.com
Ning Chen is the main contributor.
[1]School of Information Science and Engineering, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China
Full list of author information is available at the end of the article

the lower pitch-frequency cepstral coefficients were discarded and the remaining coefficients were projected onto chroma bins to obtain the Chroma DCT-Reduced log Pitch descriptor. This descriptor achieved a high degree of timbre invariance and, hence, outperformed conventional PCP in the context of music matching and retrieval applications. In [13], to describe the similarity of singing voice between cover versions of popular songs, two concepts from psychoacoustics (time-varying loudness contour and critical band) were combined with conventional PCP descriptors organically to obtain Cochlear PCP (CPCP). Besides harmonic progression, melody evolution can also be used for the CSI task, for example, in [17–19], the main melody (denoted as MLD in this paper) was extracted for cover song retrieval. Recently, timbre-based descriptors are studied for the CSI task [12, 20]. In [12], a new descriptor, Modified Perceptual Linear Prediction Lifted Cepstrum (MPLPLC), was obtained by modifying the Perceptual Linear Prediction (PLP) model in automatic speech recognition field through introducing new research achievements in psychophysics and taking the difference between speech and music into consideration to make it suitable for music signal analysis. In addition, different kinds of similarity functions, such as Cross-Correlation (CC) [9], Dynamic Time Warping (DTW) [11], Qmax [5], and Dmax [21], have been proposed for measuring the similarity between descriptors.

However, since different descriptors are based on various assumptions, designed for representing distinct characteristics of music, and often differ in scale and noise level, it is impossible to characterize all songs of different genres with the same descriptor comprehensively, nor it is possible to use only one similarity function to measure the similarity between descriptors reliably. To solve this problem, some researchers began to study descriptor or similarity fusion models for the CSI task [22–26] (see Section 2). In this paper, we propose a two-layer similarity fusion model for the CSI task aiming at enhancing the identification accuracy and classification efficiency further. The main contributions of this paper include (i) our model, combining the advantages of two musical descriptors and two similarity functions, generates more comprehensive and reliable similarity description between songs. (ii) The sparse group LASSO algorithm [27] is included in the proposed model to select the most suitable learning method for the late fusion stage to ensure the fusion efficiency and reduce the computational complexity as well. (iii) By incorporating the advantages of early fusion and late fusion organically, the proposed model outperforms state-of-the-art fusion-based CSI schemes in identification accuracy and classification efficiency. (iv) Through projecting the ordinary similarity to probability-based similarity in late fusion, the proposed model is flexible and generic enough to include more musical descriptors and similarity functions. (v) Extensive experiments have been conducted on a music collection that is composed of 3364 samples provided by SecondHandSong (SHS)[1] to verify the efficiency of the proposed model in comparison with other CSI schemes with or without similarity fusion.

2 Information fusion for CSI

Information fusion consists of combining information originating from several sources in order to improve decision making [28]. This technique is rather commonly adopted in content-based MIR field. For instance, in [29], different descriptors were combined to improve genre classification accuracy. For the CSI task, information fusion should be a suitable idea because it is easier to capture the tonal similarity between tracks by different kinds of descriptors and similarity functions. In fact, some recent studies have suggested that version detection can be improved through the combination of different descriptors [30] or different similarity functions [24–26, 31, 32]. Generally, the information fusion in the CSI field can be performed in four levels: feature, descriptor, similarity, and decision.

The feature-level fusion is the simplest way of fusion. In [30], frames from the same moment in time were taken for both chroma and melody, and then, they were combined by creating a tuple of note or chord. Finally, to reduce the number of tuples, four different representations were proposed with different alphabet sizes. However, fusion at this level may not get desired results in practice because (i) independent analysis of different features often lead to inconsistent conclusion that are hard to integrate (for example, two tracks may be judged as the reference/cover pair by one feature and the reference/non-cover pair by another feature) and (ii) preselecting a set of features leads to biased analysis. So, in [30], the improvement achieved by the fusion scheme is limited in contrast to that achieved by chord-based representations.

The descriptor-level fusion is the strategy that combines different descriptors into one descriptor vector. The simplest way is to concatenate or merge descriptors. In [30], the chord descriptor and the melody descriptor were fused by concatenating or merging them. The problems resulted from this kind of fusion include the following: First, when performing descriptor binding of different nature/domains, normalization techniques should be applied first to standardize all descriptor values in the same range, which has been a great challenge for the machine learning community [33]. Second, concatenation or merging may result in the "curse of dimensionality" problem, which means the dimension of the descriptor space increases in such a way that the available training instances become indistinguishable and not enough for allowing the definition of a good decision hyperplane [33]. Third, concatenation or merging further dilutes the

already low signal-to-noise ratio in each descriptor. So, as shown in [30], this kind of fusion may not result in satisfactory results.

The similarity level fusion is based on the strategy known as mixture of experts. The similarity between two tracks is obtained by calculating individual pairwise distance for each descriptor and then combining them into a final pairwise distance value. Several similarity level fusion schemes have been proposed for the CSI task. In [22], the main melody and accompaniment of the music were extracted first. Then, the maximum value of the similarities based on main melody, accompaniment, and mixture signal separately was taken as the final similarity. In [24], the task of detecting cover versions was posed as a classification problem. The similarities based on different descriptors and corresponding similarity functions were concatenated as a feature vector, which was then used to train a classifier for determining whether the corresponding two tracks belong to the reference/cover or reference/non-cover pair. Since only chroma descriptors were considered, the fused similarity only accounted for the same musical facet, the harmony. To solve this problem, in [25], the similarities based on three related yet different descriptors (harmony, melody, and bass line) were fused with the power of a standard classification approach similar to [24]. In [31], the fusion of different similarities was achieved by projecting all similarities in a multi-dimensional space, where the dimensionality of this space was the number of similarities considered. In [26], the similarities based on different descriptors and corresponding similarity functions were obtained first. Then, the Similarity Network Fusion (SNF) technique [34] was used to fuse the similarity communities based on each similarity. Finally, the track-by-track similarities in the fused similarity network were adopted for version identification. Due to the merits of SNF technique, this fusion scheme could reduce the noise existing in each similarity network and integrate the common as well as complementary information caught by different descriptors and corresponding similarity measures.

Finally, the fourth strategy for fusion is known as decision-level fusion. The CSI scheme proposed in [35] belongs to this kind of fusion. First, the similarities based on different descriptors were adopted to train the classifier. Then, the decision made by each classifier was integrated with standard rank aggregation. This fused scheme achieved an increase of up to 23.5% identification accuracy compared to single classifiers.

According to the stage at which the fusion is performed, the above fusion schemes can be classified into early fusion and late fusion. Early fusion happens before the classification step. The feature-level, the descriptor-level, and the similarity-level fusions belong to early fusion. The main advantage of early fusion is that all the features can be "seen" by the classifier and only one learning phase is required [36]. However, the performance of early fusion will be greatly affected by including features of little contribution. On the other hand, the late fusion approach operates at the decision level [37]. When compared to early fusion, late fusion is easier to perform, but it cannot learn the correlation among features. Usually another learning procedure is needed to combine these classification outputs. To avoid the over-fitting problem, simple mean, which can yield better or at least comparable results as those training another classifier for fusion, can be adopted [36].

In this work, we propose a two-level fusion model, which integrates the advantages of both early fusion and late fusion organically, for the CSI task. Concretely, in the early fusion, the similarities based on a specific musical descriptor (HPCP [15] or MLD [38]) by two different similarity functions (Qmax [5] and Dmax [21]) are fused with the SNF technique. In the late fusion, one optimal classifier, which is selected by the sparse group LASSO technique [27], is performed on each early fused similarity to obtain the probability that the corresponding tracks belong to the reference/cover pair. Then, the mean value of these probabilities are obtained as the final fused similarity.

3 Proposed CSI scheme

The block diagram of the proposed scheme is shown in Fig. 1.

3.1 Similarity calculation: Qmax and Dmax

The Qmax [16] similarity measure tries to calculate the length of the longest time segment in which two sequences $\mathbf{f}_i$ and $\mathbf{f}_j$ exhibit similar patterns. First, a Cross Recurrence Plot (CRP), denoted as $\mathbf{c}$, is generated by setting its element $c_{p,q}$ to "1" when there exists recurrence between $f_i(p)$ and $f_j(q)$ and "0" otherwise. More details about the CRP calculation can be found in [5]. In the CRP, the length of the diagonal pattern of "1" indicates the degree of similarity between these two sequences. However, as shown in [32], due to the possible alignment constraints in the Qmax (see Fig. 2a), it fails to identify the cover versions when the CRP includes such phenomenon as shown in Fig. 3a, where there is serious short disruption of diagonal. This phenomenon may be resulted from the skip of some chords or part of the melody when performing the cover version. To solve this problem, we modified the Qmax by changing the possible alignment constraints from Fig. 2a, b to obtain a new measure, called Dmax [21]. As shown in Fig. 3b, c, for the case shown in Fig. 3a, the Dmax performs better than the Qmax.

In the Qmax and Dmax measures, first, a cumulative matrix (denoted as $\mathbf{o}$ and $\hat{\mathbf{o}}$, respectively) is generated based on $\mathbf{c}$ with Eqs. (1) and (2), respectively.

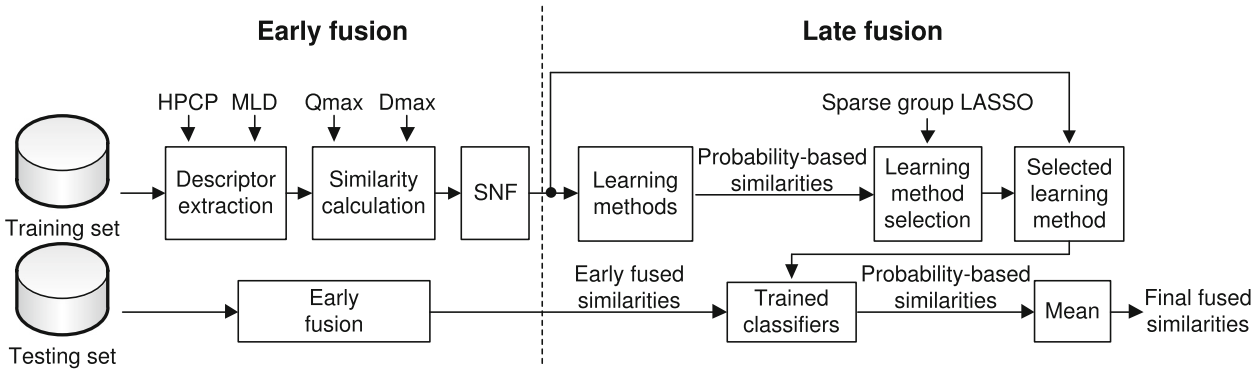

Fig. 1 Block diagram of the proposed CSI scheme

$$o_{p,q} = \begin{cases} \max\{o_{p-1,q-1}, o_{p-2,q-1}, o_{p-1,q-2}\} + 1, & if\ c_{p,q} = 1 \\ \max\{0, o_{p-1,q-1} - \gamma(c_{p-1,q-1}), \\ o_{p-2,q-1} - \gamma(c_{p-2,q-1}), \\ o_{p-1,q-2} - \gamma(c_{p-1,q-2})\}, & if\ c_{p,q} = 0 \end{cases} \tag{1}$$

$$\hat{o}_{p,q} = \begin{cases} \max\{\hat{o}_{p-1,q-1}, \hat{o}_{p-2,q-1} + c_{p-1,q}, \\ \hat{o}_{p-1,q-2} + c_{p,q-1}, \\ \hat{o}_{p-3,q-1} + c_{p-2,q} + c_{p-1,q}, \\ \hat{o}_{p-1,q-3} + c_{p,q-2} + c_{p,q-1}\} + 1, & if\ c_{p,q} = 1 \\ \max\{0, \hat{o}_{p-1,q-1} - \gamma(c_{p-1,q-1}), \\ \hat{o}_{p-2,q-1} + c_{p-1,q} - \gamma(c_{p-2,q-1}), \\ \hat{o}_{p-1,q-2} + c_{p,q-1} - \gamma(c_{p-1,q-2}), \\ \hat{o}_{p-3,q-1} + c_{p-2,q} + c_{p-1,q} - \gamma(c_{p-3,q-1}), \\ \hat{o}_{p-1,q-3} + c_{p,q-2} + c_{p,q-1} - \gamma(c_{p-1,q-3})\}, & if\ c_{p,q} = 0 \end{cases} \tag{2}$$

In both Eqs. (1) and (2), γ is calculated with Eq. (3).

$$\gamma(z) = \begin{cases} \gamma_o, & if\ z = 1 \\ \gamma_e, & if\ z = 0 \end{cases} \tag{3}$$

where γ_o and γ_e are the penalty for a disruption onset and a disruption extension, respectively.

Then, the normalized *Qmax* distance and *Dmax* distance, denoted as $d_Q(i,j)$ and $d_D(i,j)$, can be respectively calculated with Eqs. (4) and (5).

$$d_Q(i,j) = \sqrt{N_j} / \max(o_{p,q}) \tag{4}$$

$$d_D(i,j) = \sqrt{N_j} / \max(\hat{o}_{p,q}) \tag{5}$$

where N_j is the length of $\mathbf{f}_j$.

Suppose the track collection is composed of N tracks and $\mathbf{f}_i^{(k)}, k = 1, \cdots, K$ is the k-th kind of descriptor of i-th track. For $\mathbf{f}_i^{(k)}$ and $\mathbf{f}_j^{(k)}$, their similarities based on the Qmax function and Dmax function are denoted as $d_Q^{(k)}(i,j)$ and $d_D^{(k)}(i,j)$, respectively.

3.2 Early fusion: SNF

The early fusion is realized by the SNF technique [34]. With any music descriptor, the track similarity networks based on Qmax and Dmax are represented as graphs $G_Q\left(V, E_Q\right)$ and $G_D\left(V, E_D\right)$, respectively. The vertices V correspond to the track collection, and the edges E_Q (or E_D) are weighted by similarity based on Qmax (or Dmax). To compute the fused similarity matrix from the Qmax and Dmax matrices, the full kernels (denoted as

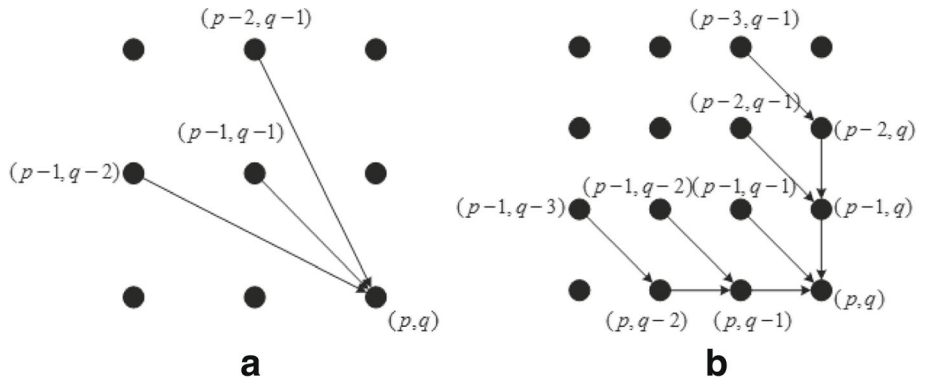

Fig. 2 Possible alignment constraints in the **a** Qmax and **b** Dmax

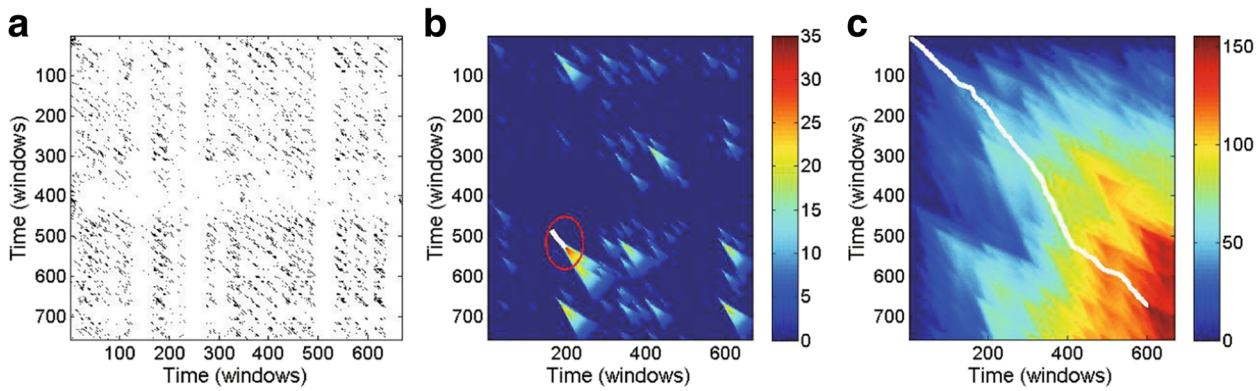

Fig. 3 a The CRP for the song "Addicted to Love" as performed by Tina Turner and Robert Palmer and the corresponding cumulative matrix obtained by the **b** Qmax and **c** Dmax

$\mathbf{P}_Q$ and $\mathbf{P}_D$) and the sparse kernels (denoted as $\mathbf{Q}_Q$ and $\mathbf{Q}_D$) are defined on the vertex set V (see Eqs. (6)–(9)), respectively.

$$P_Q(i,j) = \begin{cases} \frac{d_Q(i,j)}{2\sum_{k\neq i} d_Q(i,k)}, & j \neq i \\ 1/2, & j = i \end{cases} \tag{6}$$

$$P_D(i,j) = \begin{cases} \frac{d_D(i,j)}{2\sum_{k\neq i} d_D(i,k)}, & j \neq i \\ 1/2, & j = i \end{cases} \tag{7}$$

Let $N_{i,Q}$ (or $N_{i,D}$) represent a set of i-th track's neighbors including itself in G_Q (or G_D). For the given graph G_Q (or G_D), the K Nearest Neighbors (KNN) is used to measure local affinity as Eq. (8) (or Eq. (9)).

$$Q_Q(i,j) = \begin{cases} \frac{d_Q(i,j)}{\sum_{k\in N_{i,Q}} d_Q(i,k)}, & j \in N_{i,Q} \\ 0, & otherwise \end{cases} \tag{8}$$

$$Q_D(i,j) = \begin{cases} \frac{d_D(i,j)}{\sum_{k\in N_{i,D}} d_D(i,k)}, & j \in N_{i,D} \\ 0, & otherwise \end{cases} \tag{9}$$

Let $\mathbf{P}_{Q,t=0} = \mathbf{P}_Q$ be the initial status matrix at $t = 0$. The similarity matrix based on Qmax measure is iteratively updated with Eq. (10). After each iteration, normalization (Eq. (8)) is performed on $\mathbf{P}_{Q,t+1}$. $\mathbf{P}_{D,t+1}$ is obtained by the same way.

$$\mathbf{P}_{Q,t+1} = \mathbf{Q}_Q \times \left(\mathbf{P}_{D,t}\right) \times \left(\mathbf{Q}_Q\right)^T \tag{10}$$

After t steps, the overall status matrix, denoted as $\mathbf{P}$, is obtained with Eq. (11).

$$\mathbf{P} = \left(\mathbf{P}_{Q,t} + \mathbf{P}_{D,t}\right)/2 \tag{11}$$

For the k-th descriptor, $\mathbf{f}^{(k)}$, the corresponding fused similarity network is denoted as $G^{(k)}(V, E^{(k)})$. The weights of each edge in $G^{(k)}(V, E^{(k)})$ are concatenated to generate the k-th early fused similarity vector, denoted as $\mathbf{X}^{(k)} = [X_1^{(k)}, \cdots, X_{N^2}^{(k)}]^T$.

3.3 Late fusion and group LASSO algorithm

First, for each descriptor, different learning methods, denoted as $\mathbf{U} = \{U_1, \cdots, U_M\}$, are performed on the early fused similarities $X_n^{(k)}, n = 1, \cdots, N^2$. For $X_n^{(k)}$, the probability that it belongs to the reference/cover pair obtained by different learning methods are concatenated together to obtain $\hat{\mathbf{X}}_n^{(k)} = [\hat{X}_{1n}^{(k)}, \cdots, \hat{X}_{Mn}^{(k)}]$. Then, $\hat{\mathbf{X}}_n^{(k)}, n = 1, \cdots, N^2$ combined with their labels (reference/cover or reference/non-cover) are used to train the group LASSO [27] algorithm to select the most efficient learning method for $\mathbf{X}^{(k)}$. It should be noted that for HPCP or MLD descriptor, the early fused similarities are used to train almost all learning methods provided by Weka with default parameters, respectively. Only BayesNet (BN), NaiveBayesUpdateable (NBU), RBFNetwork (RBFN), DecisionTable (DT), and J48 methods yield good results. So, the group LASSO algorithm is then applied to the results obtained by each of these five learning methods to select the most efficient one for each descriptor. For each kind of descriptor, the probability obtained by each learning method is regarded as one group. Finally, assuming that for early fused similarities $X_n^{(\mathrm{HPCP})}$ and $X_n^{(\mathrm{MLD})}$, the corresponding probability-based similarities obtained by the most efficient learning method are $\tilde{X}_n^{(\mathrm{HPCP})}$ and $\tilde{X}_n^{(\mathrm{MLD})}$, respectively; the final fused similarity is obtained by taking the mean of $\tilde{X}_n^{(\mathrm{HPCP})}$ and $\tilde{X}_n^{(\mathrm{MLD})}$.

As shown in [39], the idea of group LASSO is to incorporate a mixed-norm regularization on logistic regression. It solves the optimization problem shown in Eq. (12).

$$\hat{\beta}_\lambda = \arg\min_{\beta,\alpha} \sum_i \log\left(1+\exp\left(-y_i\left(\beta^T\mathbf{x}_i+\alpha\right)\right)\right) + \lambda\sum_{g=1}^{G}\|\beta_{\mathbf{I}_g}\|_2 \tag{12}$$

where $\mathbf{x}_i$ is the ith training sample, y_i is the ground truth label ($\{0,1\}$), and α is the intercept. $\|\cdot\|_2$ refers to the ℓ_2 norm. β is composed of G predefined non-overlapping groups, and $\mathbf{I}_g$ is the index set of the gth group. Parameter λ controls the level of sparsity of the resulting model.

To select the most efficient learning method for the early fused similarities, the results of each learning methods (the probabilities that the similarities belong to the reference/cover pairs) are concatenated together to form the vector $\mathbf{x}_i$ in Eq. (12). Then, for a fixed λ, Eq. (12) is solved to get $\hat{\beta}_\lambda$. The learning method corresponding to the largest $\hat{\beta}_\lambda$ value is considered as the most efficient one.

4 Experiments

All the descriptors, similarity functions, and learning methods adopted in this work are listed in Table 1.

4.1 Datasets

The dataset, denoted as DB3364, is composed of tracks included in the test set of the SHS dataset. There are 1212 original tracks and 2152 cover versions. All the audio files are obtained by songs on our own. The average number of tracks in each cover set is 2.76, ranging from 2 to 17[2]. Furthermore, DB3364 is split into one training set, denoted as DB801, and three testing sets, denoted as DB799, DB802, and DB962, respectively. These datasets are not overlapping, and their specific information is listed in Table 2. It should be noted that we did not use the descriptors provided by the SHS dataset directly. The HPCP, MLD, CPCP, and Beat Synchronous Chroma (BSC) descriptors were extracted from the audio with the algorithms shown in [13, 15, 19], and [9], respectively.

Table 1 The descriptors, similarity functions, and learning methods used

	Abbreviations	Descriptions
Descriptors	HPCP	Harmonic Pitch Class Profiles [15]
	CPCP	Cochlear Pitch Class Profiles [13]
	MLD	Melody [19]
	BSC	Beat-Synchronous Chroma [9]
Similarity functions	Dmax	The Dmax similarity measure [21]
	Qmax	The Qmax similarity measure [5]
Learning methods	BN	BayesNet
	RBFN	RBFNetwork
	J48	J48
	DT	DecisionTable
	NBU	NaiveBayesUpdateable

Table 2 Cover song datasets used

Datasets	Num of tracks	Num of cover sets	Ave. num of tracks in each cover set
DB801	801	273	2.93
DB799	799	283	2.82
DB802	802	279	2.87
DB962	962	377	2.55

4.2 Evaluation measures

With the final similarity obtained by each CSI scheme, an ordered list of results for each given query can be obtained. Then, the identification accuracy can be evaluated using standard information retrieval metrics, the mean of average precision (MAP)[5], Mean averaged reciprocal rank (MaRR)[40], total number of covers identified in top 10 (TOP10), and mean rank (MR) of first correctly identified cover.

In addition, the final fused similarities are adopted to train a classifier (BayesNet classifier provided by Weka[3] with default parameters), which can then be used to estimate whether any two tracks belong to the reference/cover pair or not. Assume the obtained confusion matrix is as Table 3, where class A and class B denote the reference/non-cover and reference/cover class, respectively. Then, three different parameters, True Negative Rate (TNR), Classification Accuracy (CA), and Average Classification Accuracy (ACA), are calculated according to Eqs. (13)–(15) to measure the classification efficiency. The ACA is adopted to avoid that the evaluation of classification results are biased towards the majority class (the reference/non-cover class in this work). Since 10-fold cross-validation protocol is adopted, all reported results are in terms of mean TNR, mean CA, and mean ACA.

$$\mathrm{TNR} = \mathrm{TN}/(\mathrm{FN}+\mathrm{TN}) \tag{13}$$

Table 3 Confusion matrix

		Predicted	
		Class A	Class B
Real	Class A	TP	FP
	Class B	FN	TN

$$\mathrm{CA} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}) \quad (14)$$

$$\mathrm{ACA} = (\mathrm{TN}/(\mathrm{FP} + \mathrm{TN}) + \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}))/2 \quad (15)$$

4.3 Experimental results

To illustrate how the proposed model behaves in easy and hard conditions, we manually choose concrete cover sets where one descriptor performs better than the other and where both descriptors perform well. The information of the tracks included in this study is listed in Table 4. The six tracks are used both as the queries and the targets. The corresponding 6×6 distance matrices obtained by HPCP-Dmax, HPCP-Qmax, 1L-HPCP-QD (early fusion based on HPCP descriptor), MLD-Dmax, MLD-Qmax, 1L-MLD-QD (early fusion based on MLD descriptor), 2L-HPCP-Best1 (two-layer fusion when the BayesNet classifier is only applied on HPCP-QD similarity), 2L-MLD-Best1 (two layer fusion when the BayesNet classifier is only applied on MLD-QD similarity), and 2L-Best1 (two-layer fusion when the BayesNet classifier is applied on both HPCP-QD and MLD-QD similarities) are shown in Fig. 4a–i, respectively. The cells corresponding to the query/cover pairs are marked with blue boxes.

The experimental results shown in Fig. 4 demonstrate that (i) by comparing the results shown in Fig. 4a, b, d, e, we can see that the HPCP descriptor works better than the MLD descriptor for No. 1 cover set, the MLD descriptor performs better than the HPCP descriptor for No. 3 cover set, and both HPCP and MLD descriptors perform well on No. 2 cover set. The tracks in the No. 1 cover set are two different versions of "Something Wonderful". These two tracks are mainly composed of the sound of stringed and wind instruments, and they include no prominent melody. In this case, the HPCP descriptor, which can describe the harmonic progression very well, performs better than MLD descriptor. The No. 3 cover set is composed of two versions of "Never Can Say Goodbye" performed by different singers. Both of them include main melody performed by female singer, and the accompaniment in these two tracks is weak when compared with the vocal sound. In this circumstance, the MLD descriptor performs better than the HPCP descriptor. Since the two tracks in No. 2 cover set include strong accompaniment and predominant melody, both the HPCP and MLD descriptors perform well on it. (ii) By comparing the results among Fig. 4a–c and those among Fig. 4d–f, we can see that when the single similarity (HPCP-Dmax in this case) can not distinguish the reference/cover pair (No. 3 cover set in this case), the early fused similarity (1L-HPCP-QD in this case) can perform very well, and when the two single similarity measures (MLD-Dmax and MLD-Qmax in this case) perform well on No. 1 cover set, the early fused similarity (1L-MLD-QD in this case) performs better on it. The latent reason is that the early fusion method may utilize the complementarity between the Qmax and Dmax in finding alignments. (iii) As shown in Fig. 4g–i, when compared with 2L-HPCP-Best1 or 2L-MLD-Best1-based schemes, 2L-Best1 achieves global optimum on three cover sets. The possible reason is that the late fusion method may fuse the complementarity between the two early fused similarities (1L-HPCP-QD and 1L-MLD-QD in this case) efficiently. (iv) By comparing the results in Fig. 4 c, f, i, we can see that when compared with the early fusion methods (1L-HPCP-QD and 1L-MLD-QD in this case), the late fusion may enlarge the distance between the intro-distance and inter-distance further, which may result in a higher identification accuracy.

Table 4 The tracks in the cover sets

Cover sets	Title of the tracks	Track ID	Artists
No. 1	Something Wonderful	1	Carly Simon
		2	Amel Larrieux
No. 2	Spooky	3	The Puppini Sister
		4	Atlanta Rhythm Section
No. 3	Never Can Say Goodbye	5	Isaac Hayes
		6	Gerald Albright

4.3.1 Efficiency of early fusion

To test the validity of the early fusion, the identification accuracy (in terms of MAP, MaRR, TOP10, and MR) and classification efficiency (in terms of TNR, CA, and ACA) obtained before and after the early fusion are compared in Fig. 5, where HPCP+QD [32] (or MLD+QD, CPCP+QD, and BSC+QD) denotes the early fused similarity based on HPCP+Qmax [5] (or MLD+Qmax, CPCP+Qmax, and BSC+Qmax) measure and HPCP+Dmax (or MLD+Dmax, CPCP+Dmax, and BSC+Dmax) measure. We observe that the performances (in terms of all the evaluation measures except for MR) achieved by the early fusion scheme are much better than those of the fused objects (including the scheme proposed in [5]), which verifies that these two similarity measures (Qmax and Dmax) carry complementary information. As shown in [34], the early fusion scheme can integrate this information efficiently because (i) the low-weight edges in each similarity network are cut, which helps to reduce the noise and (ii) the high-weight edges present in one or two networks are added to the other and the low-weight edges supported by both networks are retained depending on how tightly connected their neighborhoods are across networks, which helps to integrate common as

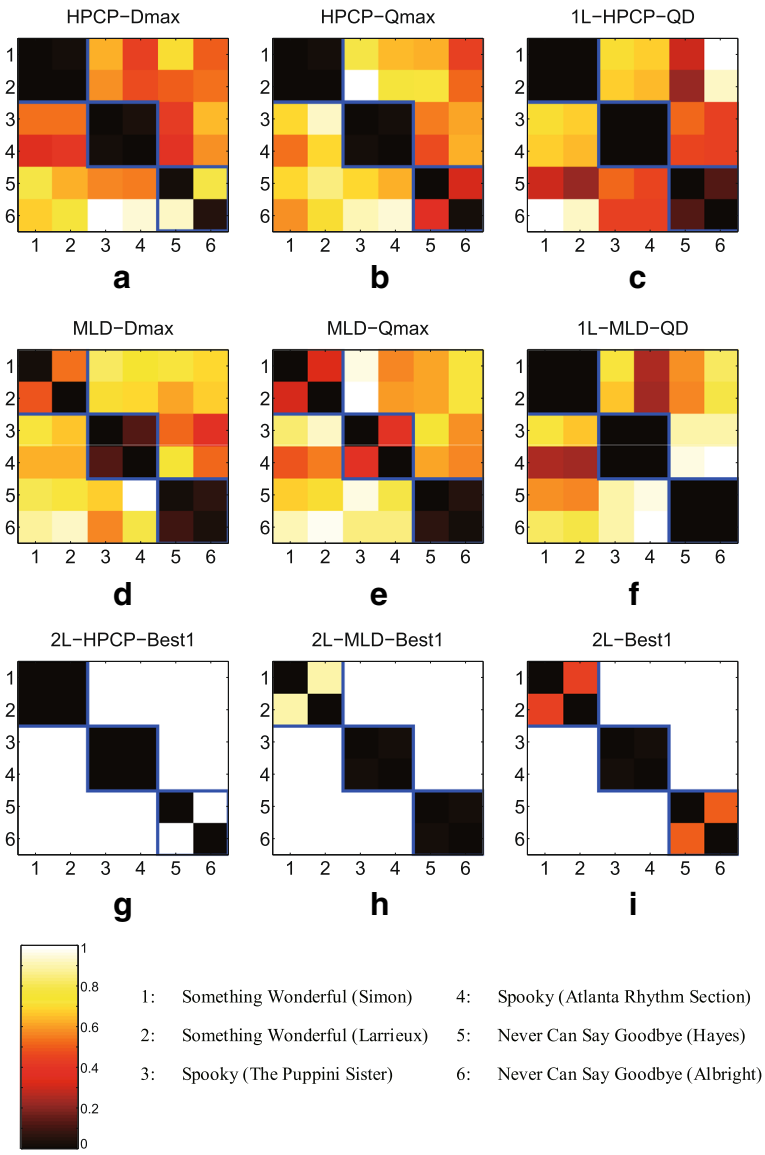

Fig. 4 Distance matrices obtained by CSI schemes with or without similarity fusion. The actual values are subtracted by 1 to make the visual comparison easier. **a** HPCP-Dmax. **b** HPCP-Qmax. **c** 1L-HPCP-QD. **d** MLD-Dmax. **e** MLD-Qmax. **f** 1L-MLD-QD. **g** 2L-HPCP-Best1. **h** 2L-MLD-Best1. **i** 2L-Best1

well as complementary information across the similarity networks.

4.3.2 Learning method selection

The averages of the magnitudes of regression coefficients obtained by the group LASSO for each learning method and descriptor considered are plotted in Fig. 6. It can be seen that across the learning methods we choose, the BayesNet is the most efficient for both HPCP and MLD.

In addition, we compare the performances obtained after late fusion under different learning method combinations in Fig. 7, where 2L-Best1 denotes the combination of the classification results of the top one learning method for HPCP+QD similarity and that for MLD+QD similarity, the 2L-Best2 means the combination of the classification results of the top two learning method for HPCP+QD similarity and those for MLD+QD similarity, and so on. We observe that (i) different learning method combinations obtain the similar MAP and MaRR performances on all four datasets and 2L-Best1 performs more stably across different datasets than the other combinations. (ii) 2L-Best1 achieves consistently better performances, in terms of TOP10, TNR, and

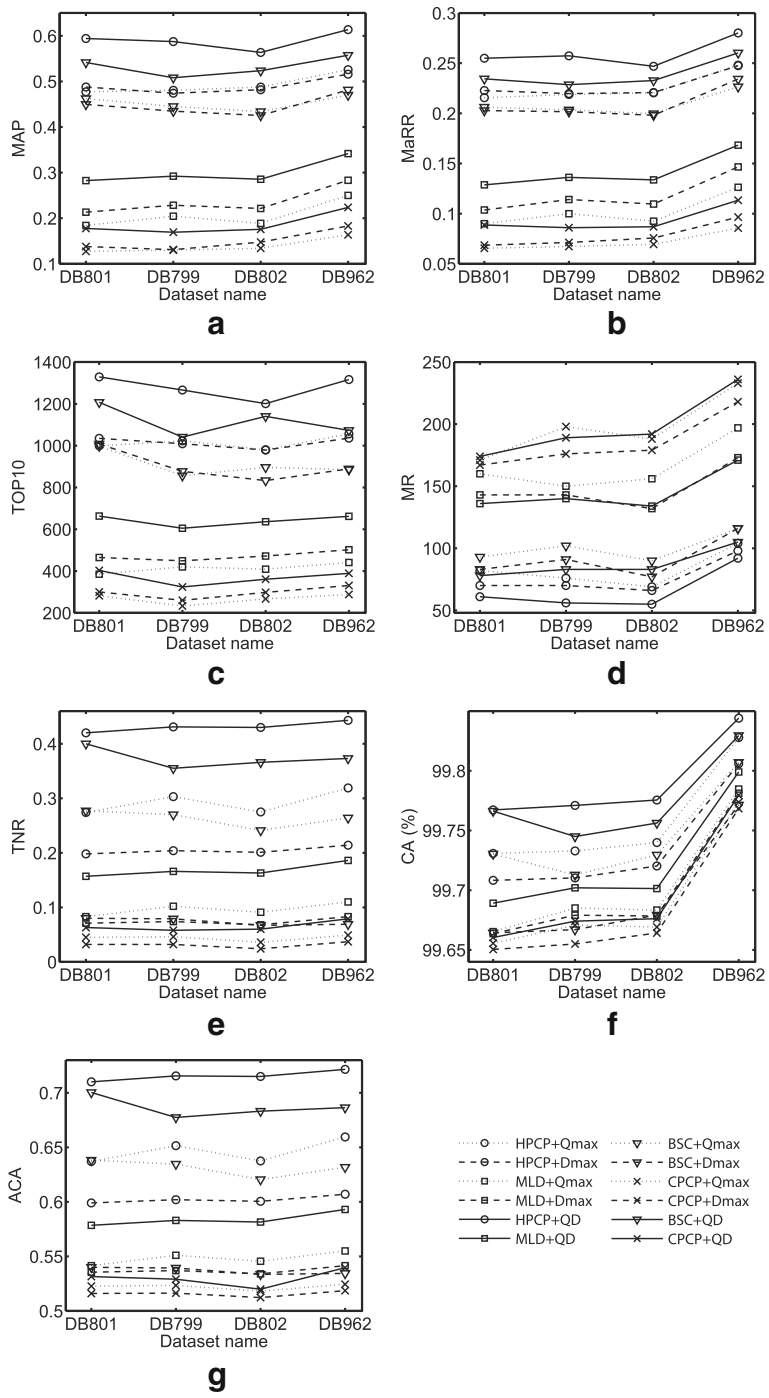

Fig. 5 Comparison of the identification accuracy in terms of **a** MAP, **b** MaRR, **c** TOP10, and **d** MR and the classification efficiency in terms of **e** TNR, **f** CA, and **g** ACA before and after early fusion on different data sets. *QD* early fusion of the Qmax and Dmax-based similarities with SNF

ACA, than the other combinations on all four datasets. (iii) In Fig. 7d, the lines with circle, triangle, and square overlap, which means that 2L-Best1 scheme performs similar as 2L-Best2 or 2L-Best3 scheme and better than 2L-Best4 or 2L-Best5 scheme in term of MR. (iv) As shown in Fig. 7f, 2L-Best1 performs worse than the other four combinations, but the gap is very small (about 0.01%). In general, 2L-Best1 combination performs much

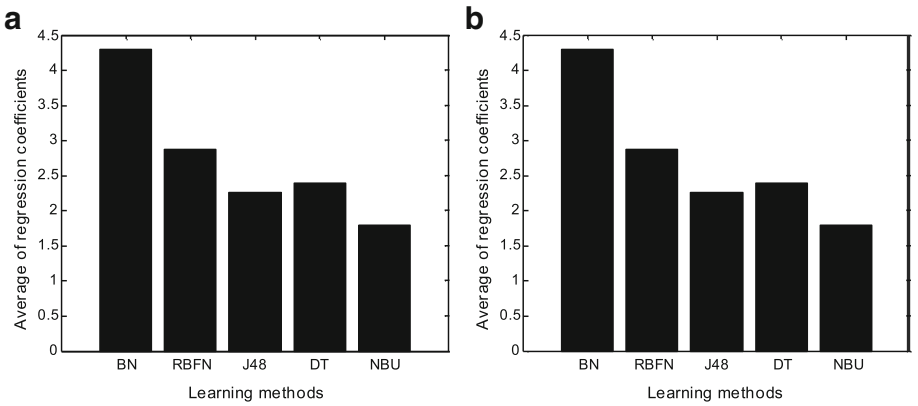

Fig. 6 Averages of the magnitudes of regression coefficients across the learning methods included for **a** HPCP and **b** MLD

better than the other four combinations especially when computational complexity is considered. So, 2L-Best1 is adopted to obtain the final fused similarity. Specifically, the mean of the probability-based similarities obtained by the BayesNet classifier for HPCP-QD similarity and for MLD-QD similarity is taken as the final fused similarity.

4.3.3 Complementarity between different descriptors

Another important question is whether the information carried by different descriptors is complementary. To answer this question, the performances obtained by 2L-HPCP-Best1 and those achieved by 2L-Best1 combination are compared in Tables 5 and 6. We observe that 2L-Best1 scheme performs better than 2L-HPCP-Best1 scheme across all the evaluation measures included on all four datasets except for the CA performance on DB802, where the gap is less than 0.005%. So it is verified that the MLD-QD and HPCP-QD similarities carry complementary information. As shown in Tables 5 and 6, the similar conclusion can also be obtained for the CPCP-QD and BSC-QD similarities. Therefore, the combination of different descriptors can help to improve performances.

4.3.4 Comparison with state-of-the-art fusion based CSI schemes

In these experiments, the performances of eight fusion techniques are compared on all four datasets. They are two-layer fusion with best learning method (2L-Best1), two-layer fusion when HPCP is considered (2L-HPCP-Best1), early fusion based on HPCP descriptor (1L-HPCP-QD) [32] and on MLD descriptor (1L-MLD-QD), the schemes proposed in [24, 26], and [31], and the Particle Swarm Optimization (PSO) based one. In PSO based scheme, the similarities used in this work, HPCP+Qmax, HPCP+Dmax, MLD+Qmax and MLD+Dmax, are weighted and added together, and the optimal weight combination is sought by PSO technique [41]. For the fusion schemes in [26] and [24][4], the fused objects are those provided in [26] and [24], respectively. Unfortunately, we can not obtain the implementation of the pitch salience function used in [31], so the fused objects for [31] are those used in this work.

The comparison results in terms of identification accuracy and classification efficiency are shown in Tables 5 and 6, respectively. It can be seen that: i) For HPCP-MLD (or BSC-CPCP) based combination, 2L-Best1 scheme performs much better than 1L-HPCP-QD [32] (or 1L-BSC-QD) or 1L-MLD-QD (or 1L-CPCP-QD) scheme in terms of all evaluation measures on all four datasets except for CA on DB802 (where the gap is smaller than 0.005%), which verifies the necessity and validity of the late fusion. ii) For HPCP-MLD based combination, 2L-Best1 scheme performs much better than state-of-the-art fusion based CSI schemes [24, 26, 31] and PSO based one in terms of identification accuracy and classification efficiency on all four datasets, except for the TNR on DB962, where PSO scheme achieves higher TNR at the sacrifice of much lower CA. iii) For BSC-CPCP based combination, 2L-Best1 scheme performs much better than the fusion based CSI scheme in [31] and PSO based one in terms of identification accuracy and classification efficiency on all four datasets. However, in some cases, 2L-Best1 scheme performs worse than the fusion based CSI schemes in [24, 26]. The possible reason is that the number and the type of the descriptors fused in [24, 26] are different from those used in BSC-CPCP based combination.

As shown in Tables 5 and 6, when the similar experiments are applied on the CPCP- and BSC-based similarities, the similar results are obtained. It should be noted that since there are too many reference/non-cover pairs in the training set, the classifier would determine that almost

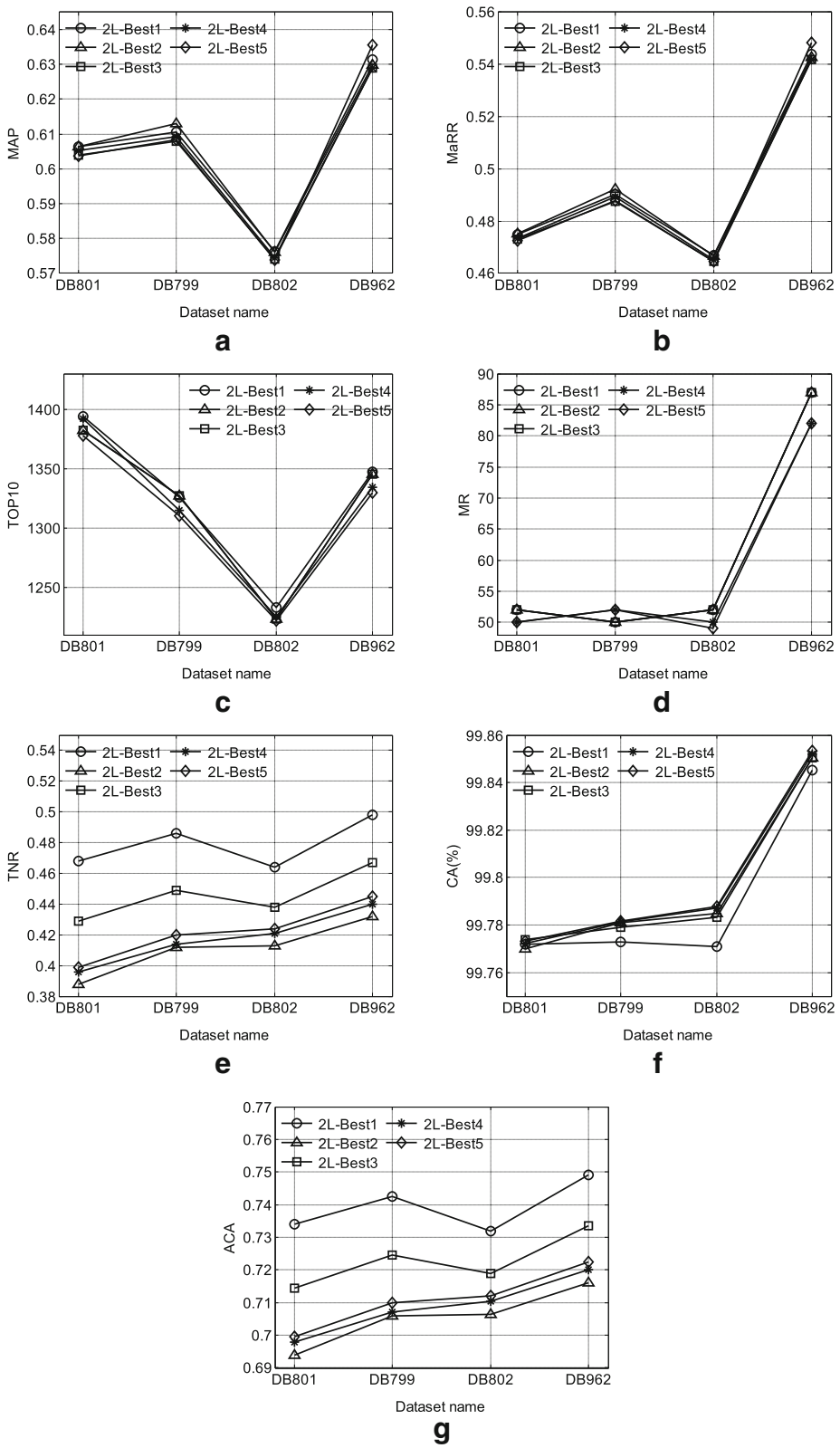

Fig. 7 Comparison of identification accuracy, in terms of **a** MAP, **b** MaRR, **c** TOP10, and **d** MeanRank, and classification efficiency, in terms of **e** TNR, **f** CA, and **g** ACA, obtained after late fusion based on different learning method combinations on different datasets

Table 5 Identification accuracy achieved by different descriptor combinations

	Combination of HPCP and MLD					Combination of BSC and CPCP				
		MAP	MaRR	TOP10	MR		MAP	MaRR	TOP10	MR
DB801	2L-Best1	0.6064	0.4748	1394	52	2L-Best1	0.5563	0.4396	1251	72
	2L-HPCP-Best1	0.5898	0.4628	1249	61	2L-BSC-Best1	0.5394	0.4255	1233	88
	1L-HPCP-QD [32]	0.5941	0.2549	1329	61	1L-BSC-QD	0.5415	0.2344	1207	78
	1L-MLD-QD	0.2825	0.1287	663	136	1L-CPCP-QD	0.1776	0.0887	403	174
	[26]	0.5393	0.2358	1138	84	[26]	0.5393	0.2358	1138	84
	[24]	–	–	–	–	[24]	–	–	–	–
	PSO	0.4899	0.2201	1030	77	PSO	0.4732	0.2115	1031	86
	[31]	0.3939	0.1842	859	77	[31]	0.3242	0.1553	763	90
DB799	2L-Best1	0.6105	0.4901	1326	50	2L-Best1	0.5280	0.4419	1105	78
	2L-HPCP-Best1	0.5824	0.4676	1268	57	2L-BSC-Best1	0.5050	0.4204	1049	90
	1L-HPCP-QD [32]	0.5873	0.2573	1266	56	1L-BSC-QD	0.5080	0.2286	1041	83
	1L-MLD-QD	0.2922	0.1362	605	140	1L-CPCP-QD	0.1693	0.086	324	189
	[26]	0.5418	0.2313	1122	78	[26]	0.5418	0.2313	1122	78
	[24]	–	–	–	–	[24]	–	–	–	–
	PSO	0.5025	0.2280	1058	70	PSO	0.4491	0.2165	8671	99
	[31]	0.4281	0.2009	894	67	[31]	0.2963	0.1460	644	96
DB802	2L-Best1	0.5764	0.4668	1233	52	2L-Best1	0.5461	0.4469	1169	67
	2L-HPCP-Best1	0.5609	0.4515	1208	66	2L-BSC-Best1	0.5166	0.4210	1143	96
	1L-HPCP-QD [32]	0.5635	0.2469	1201	55	1L-BSC-QD	0.5234	0.2327	1140	83
	1L-MLD-QD	0.2856	0.1337	636	134	1L-CPCP-QD	0.1757	0.0870	361	192
	[26]	0.5228	0.2318	1072	82	[26]	0.5228	0.2318	1072	82
	[24]	–	–	–	–	[24]	–	–	–	–
	PSO	0.5050	0.2277	1026	63	PSO	0.4401	0.2029	920	85
	[31]	0.4280	0.1986	896	61	[31]	0.3196	0.1552	728	86
DB962	2L-Best1	0.6315	0.5440	1347	87	2L-Best1	0.5833	0.5186	1124	105
	2L-HPCP-Best1	0.6125	0.5242	1334	102	2L-BSC-Best1	0.5567	0.4921	1098	129
	1L-HPCP-QD [32]	0.6135	0.2801	1316	92	1L-BSC-QD	0.5573	0.2602	1073	105
	1L-MLD-QD	0.3417	0.1683	662	174	1L-CPCP-QD	0.2238	0.1134	389	236
	[26]	0.5856	0.2710	1170	105	[26]	0.5856	0.2710	1170	105
	[24]	–	–	–	–	[24]	–	–	–	–
	PSO	0.5452	0.2568	1098	91	PSO	0.4851	0.2336	908	111
	[31]	0.4568	0.2213	917	89	[31]	0.3627	0.1819	699	113

Notes: For MAP, MaRR, and TOP10, the larger the value is, the better performance is obtained. For MR, the smaller the value is, the better performance is obtained

all reference/test pairs are reference/non-cover pairs during testing the classifier. So, almost all CA values in Table 6 are around 98 and 99%.

5 Conclusions

To date, few investigations have been done on describing similarity between versions by combining different musical descriptors and similarity functions. In this paper we take a necessary step in this field, which can not only improve performances in terms of identification accuracy and classification efficiency but also improve our understanding of the relationship between different musical descriptors and similarity functions. Two musical descriptors and two different similarity functions are fused by a two-layer fusion model. In the early fusion, the similarities obtained by applying different similarity functions on one music descriptor are fused by the SNF technique. In the late fusion, the early fused similarities based on two different musical descriptors are integrated through mapping the similarities to probability-based

Table 6 Classification efficiency achieved by different descriptor combinations

	Combination of HPCP and MLD				Combination of BSC and CPCP			
		TNR	CA (%)	ACA		TNR	CA (%)	ACA
DB801	2L-Best1	0.468	99.7718	0.7340	2L-Best1	0.417	99.7715	0.7081
	2L-HPCP-Best1	0.420	99.7692	0.7100	2L-BSC-Best1	0.401	99.7662	0.7003
	1L-HPCP-QD [32]	0.420	99.7672	0.7100	1L-BSC-QD	0.401	99.7662	0.7003
	1L-MLD-QD	0.157	99.6891	0.5785	1L-CPCP-QD	0.063	99.6604	0.5316
	[26]	0.359	99.7541	0.6795	[26]	0.359	99.7541	0.6795
	[24]	0.456	98.8105	0.7230	[24]	0.456	98.8105	0.7230
	PSO	0.279	99.7289	0.6395	PSO	0.242	99.7218	0.6208
	[31]	0.114	99.6749	0.5570	[31]	0.089	99.6689	0.5446
DB799	2L-Best1	0.486	99.7729	0.7425	2L-Best1	0.370	99.7778	0.6847
	2L-HPCP-Best1	0.431	99.7710	0.7155	2L-BSC-Best1	0.355	99.7450	0.6774
	1L-HPCP-QD [32]	0.431	99.7710	0.7155	1L-BSC-QD	0.355	99.7450	0.6774
	1L-MLD-QD	0.166	99.7020	0.5830	1L-CPCP-QD	0.058	99.6741	0.5291
	[26]	0.347	99.7541	0.6735	[26]	0.347	99.7541	0.6735
	[24]	0.408	98.7142	0.6985	[24]	0.408	98.7142	0.6985
	PSO	0.259	99.7360	0.6295	PSO	0.288	99.6281	0.6435
	[31]	0.112	99.6931	0.5560	[31]	0.066	99.6772	0.5332
DB802	2L-Best1	0.464	99.7709	0.7320	2L-Best1	0.372	99.7565	0.6859
	2L-HPCP-Best1	0.430	99.7755	0.7150	2L-BSC-Best1	0.366	99.7562	0.6832
	1L-HPCP-QD [32]	0.430	99.7755	0.7150	1L-BSC-QD	0.366	99.7562	0.6832
	1L-MLD-QD	0.163	99.7014	0.5815	1L-CPCP-QD	0.060	99.6762	0.5299
	[26]	0.340	99.7572	0.6700	[26]	0.340	99.7572	0.6700
	[24]	0.407	99.1071	0.7000	[24]	0.407	99.1071	0.7000
	PSO	0.429	98.5529	0.7080	PSO	0.202	99.7289	0.6395
	[31]	0.101	99.6908	0.5505	[31]	0.060	99.6771	0.5299
DB962	2L-Best1	0.498	99.8453	0.7490	2L-Best1	0.383	99.8302	0.6915
	2L-HPCP-Best1	0.443	99.8440	0.7215	2L-BSC-Best1	0.373	99.8295	0.6865
	1L-HPCP-QD [32]	0.443	99.8440	0.7215	1L-BSC-QD	0.373	99.8295	0.6865
	1L-MLD-QD	0.186	99.7990	0.5930	1L-CPCP-QD	0.079	99.7796	0.5397
	[26]	0.377	99.8354	0.6885	[26]	0.377	99.8354	0.6885
	[24]	0.423	99.1193	0.7080	[24]	0.423	99.1193	0.7080
	PSO	0.740	71.7942	0.7290	PSO	0.239	99.8040	0.6194
	[31]	0.140	99.7926	0.5700	[31]	0.085	99.7814	0.5424

Notes: For TNR, CA, and ACA, the larger the value is, the better performance is obtained

similarities by learning method and then taking the mean value of the results. In addition, the group LASSO technique is adopted to select the best learning method for each kind of early fused similarity before the late fusion to ensure the fusion efficiency and reduce the computational complexity as well. By incorporating the advantages of early fusion and late fusion, the proposed scheme achieves better performances in terms of identification accuracy and classification efficiency than state-of-the-art fusion-based CSI schemes. Another important advantage of the proposed model is that it is flexible and generic enough to include more musical descriptors and similarity functions to enhance the performance further. However, the disadvantage of the proposed scheme is that it achieves higher CSI identification accuracy at the cost of higher computation complexity.

For future work, considering the similarity between CSI task and Query-By-Humming (QBH) task, we plan to modify the proposed model to make it suitable for the QBH task.

Endnotes

[1] http://labrosa.ee.columbia.edu/millionsong/secondhand

[2] A complete list of tracks included in the music collection and the code of the proposed scheme can be found (http://nchenecust.com/download.html).

[3] http://weka.wikispaces.com/

[4] http://infiniteseriousness.weebly.com/cover-song-detection.html

Abbreviations

ACA: Average classification accuracy; BN: BayesNet; BSC: Beat-synchronous chroma; CA: Classification accuracy; CC: Cross-correlation; CPCP: Cochlear PCP; CRP: Cross recurrence plot; CSI: Cover Song Identification; DT: DecisionTable; DTW: Dynamic time warping; HPCP: Harmonic PCP; MAP: Mean of average precision; MaRR: Mean averaged reciprocal rank; MIR: Music Information Retrieval; MIREX: Music information retrieval evaluation eXchange; MLD: Melody; MPLPLC: Modified perceptual linear prediction lifted cepstrum; MR: Mean rank; NBU: NaiveBayesUpdateable; PCP: Pitch class profile; PLP: Perceptual linear prediction; PSO: Particle swarm optimization; QBH: Query-By-Humming; QD: Early fusion of Qmax and Dmax similarities based on SNF; RBEU: RBFNetwork; SHS: SecondHandSongs; SNF: Similarity network fusion; TNR: True negative rate; TOP10: total number of covers identified in top 10

Acknowledgements

This work was supported by the National Natural Science Foundation of China [grant number 61271349].

Availability of data and materials

A complete list of tracks included in the music collection and the code of the proposed scheme can be found at http://nchenecust.com/download.html.

Authors' contributions

NC conceived of the study; participated in the design of the work, data collection, data analysis, interpretation, and coordination; and drafted the manuscript. ML participated in the data collection, data analysis, and interpretation and helped to draft the manuscript. HX participated in the design of the work and critical revision of the article. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1] School of Information Science and Engineering, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China. [2] Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai, China.

Received: 30 September 2016 Accepted: 3 May 2017
Published online: 22 May 2017

References

1. MA Casey, R Veltkamp, M Goto, M Leman, C Rhodes, M Slaney, Content-based music information retrieval: current directions and future challenges. Proc. IEEE. **96**(4), 668–696 (2008)
2. J Serrà Julià, Identification of versions of the same musical composition by processing audio descriptions. PhD thesis, Universitat Pompeu Fabra (2011)
3. J Serrà, E Gómez, P Herrera, in *Advances in Music Information Retrieval*. Audio cover song identification and similarity: background, approaches, evaluation, and beyond (Springer, Springer-Verlag Berlin, 2010), pp. 307–332
4. T Fujishima, in *Proceedings of International Computer Music Association*. Realtime chord recognition of musical sound: a system using common lisp music (International Computer Music Association, Inc., San Francisco, 1999), pp. 464–467
5. J Serra, X Serra, RG Andrzejak, Cross recurrence quantification for cover song identification. N. J. Phys. **11**(9), 093017 (2009)
6. T-M Chang, E-T Chen, C-B Hsieh, P-C Chang, in *Proceedings of 2013 IEEE 2nd Global Conference on Consumer Electronics (GCCE)*. Cover song identification with direct chroma feature extraction from aac files (IEEE, Tokyo, Japan, 2013), pp. 55–56
7. TC Walters, DA Ross, RF Lyon, in *International Symposium on Computer Music Modeling and Retrieval*. The intervalgram: an audio feature for large-scale cover-song recognition (Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2012), pp. 197–213
8. X Chuan, in *Proceedings of 2012 International Conference on Systems and Informatics (ICSAI)*. Cover song identification using an enhanced chroma over a binary classifier based similarity measurement framework (IEEE, Yantai University, Yantai, China, 2012), pp. 2170–2176
9. DPW Ellis, Identifying 'Cover Songs' with Beat-Synchronous Chroma Features, MIREX extended abstract, 1–4 (2006). http://hdl.handle.net/10022/AC:P:13699
10. M Muller, S Ewert, Towards timbre-invariant audio features for harmony-based music. IEEE/ACM Trans. Audio, Speech, Lang. Process. **18**(3), 649–662 (2010)
11. J Serra, E Gómez, P Herrera, X Serra, Chroma binary similarity and local alignment applied to cover song identification. IEEE/ACM Trans. Audio, Speech, Lang. Process. **16**(6), 1138–1151 (2008)
12. N Chen, JS Downie, H Xiao, Y Zhu, J Zhu, in *Proc. ISMIR*. Modified perceptual linear prediction liftered cepstrum (mplplc) model for pop cover song recognition (the ATIC Research Group of the University of Malaga, Malaga, Spain, 2015)
13. N Chen, JS Downie, H-D Xiao, Y Zhu, Cochlear pitch class profile for cover song identification. Appl. Acoust. **99**, 92–96 (2015)
14. JS Downie, The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. Acoust. Sci. Technol. **29**(4), 247–255 (2008)
15. E Gómez, Tonal description of music audio signals. PhD thesis, Universitat Pompeu Fabra (2006)
16. Serrà, M Zanin, RG Andrzejak, in *Proceedings of 2009 International Society for Music Information Retrieval*. Cover song retrieval by cross recurrence quantification and unsupervised set detection, (Kobe, 2009), pp. 1–3
17. W-H Tsai, H-M Yu, H-M Wang, Using the similarity of main melodies to identify cover versions of popular songs for music document retrieval. J. Inform. Sci. Eng. **24**(6), 1669–1687 (2008)
18. M Marolt, A mid-level representation for melody-based retrieval in audio collections. IEEE Trans. Multimedia. **10**(8), 1617–1625 (2008)
19. J Salamon, J Serra, E Gómez, Tonal representations for music retrieval: from version identification to query-by-humming. Int. J. Multimedia Inform. Retriev. **2**(1), 45–58 (2013)
20. CJ Tralie, P Bendich, in *Proceedings of 16th International Society for Music Information Retrieval Conference*. Cover song identification with timbral shape sequences, (Malaga, Spain, 2015), pp. 38–44
21. F Yang, N Chen, Cover song identification based on cross recurrence plot and local alignment. J. East China Univ. Sci. Technol. **42**(2), 247–253 (2016)
22. R Foucard, J-L Durrieu, M Lagrange, G Richard, in *Proceedings of 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP 2010)*. Multimodal similarity between musical streams for cover version detection (IEEE, Dallas, 2010), pp. 5514–5517
23. CCS Liem, A Hanjalic, in *Proceedings of 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*. Cover song retrieval: a comparative study of system component choices, (Kobe, 2009), pp. 573–578
24. S Ravuri, DP Ellis, in *Proceedings of 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP 2010)*. Cover song detection: from high scores to general classification (IEEE, Dallas, 2010), pp. 65–68
25. J Salamon, Serrà, E Gómez, in *Proceedings of the 21st International Conference Companion on World Wide Web*. Melody, bass line, and harmony representations for music version identification (ACM, New York, 2012), pp. 887–894
26. N Chen, H-d Xiao, Similarity fusion scheme for cover song identification. Electron. Lett. **52**(13), 1173–1175 (2016)

27. J Friedman, T Hastie, R Tibshirani, A note on the group lasso and a sparse group lasso. arXiv preprint arXiv:1001.0736 (2010). https://arxiv.org/pdf/1001.0736.pdf
28. I Bloch, *Information fusion in signal and image processing: major probabilistic and non-probabilistic numerical approaches*. (John Wiley & Sons, Hoboken, 2013)
29. C McKay, I Fujinaga, in *ISMIR*. Automatic genre classification using large high-level musical feature sets, vol. 2004 (Citeseer, Audiovisual Institute, Universitat Pompeu Fabra, Barcelona, 2004), pp. 525–530
30. T Ahonen, *et al*, Cover song identification using compression-based distance measures. Series of publications A/Department of Computer Science, University of Helsinki (2016)
31. A Degani, M Dalai, R Leonardi, P Migliorati, in *Proceedings of 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2013)*. A heuristic for distance fusion in cover song identification (IEEE, Telecom ParisTech, 46 rue Barrault, Paris, 2013), pp. 1–4
32. N Chen, Cover song identification based on similarity fusion. MIREX extended abstract (2016). http://www.music-ir.org/mirex/abstracts/2016/CL1.pdf
33. FA Faria, JA Dos Santos, A Rocha, Torres RdS, A framework for selection and fusion of pattern classifiers in multimedia recognition. Pattern Recognit. Lett. **39**, 52–64 (2014)
34. B Wang, AM Mezlini, F Demir, M Fiume, Z Tu, M Brudno, B Haibe-Kains, A Goldenberg, Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. **11**(3), 333–337 (2014)
35. J Osmalsky, J-J Embrechts, P Foster, S Dixon, in *16th International Society for Music Information Retrieval Conference*. Combining features for cover song identification (the ATIC Research Group of the University of Malaga, Malaga, 2015)
36. Z-z Lan, L Bao, S-I Yu, W Liu, AG Hauptmann, Multimedia classification and event detection using double fusion. Multimedia Tools Appl. **71**(1), 333–347 (2014)
37. B Ionescu, J Benois-Pineau, T Piatrik, G Quénot, *Fusion in computer vision*. (Springer International Publishing, Switzerland, 2014)
38. J Salamon, E Gómez, J Bonada, in *Proc. of 14th Int. Conf. on Digital Audio Effects (DAFx-11)*. Sinusoid extraction and salience function design for predominant melody estimation, (Paris, 2011), pp. 73–80
39. Y Wang, K Han, D Wang, Exploring monaural features for classification-based speech segregation. IEEE Trans. Audio, Speech, Lang. Process. **21**(2), 270–279 (2013)
40. J Salamon, Melody extraction from polyphonic music signals. PhD thesis, Universitat Pompeu Fabra (2013)
41. Y Shi, *et al*, in *Proceedings of the 2001 Congress on Evolutionary Computation*. Particle swarm optimization: developments, applications and resources, vol. 1 (IEEE, Seoul, 2001), pp. 81–86