Int J Multimed Info Retr (2017) 6:155–166
DOI 10.1007/s13735-017-0125-z

REGULAR PAPER

Query-by-example music information retrieval by score-based genre prediction and similarity measure

Nastaran Borjian[1]

Received: 13 December 2016 / Revised: 25 March 2017 / Accepted: 17 April 2017 / Published online: 25 April 2017
© Springer-Verlag London 2017

Abstract A topic of music information retrieval (MIR) field is query-by-example (QBE), which searches a popular music dataset using a user-provided query and aims to find the target song. Since this type of MIR has been generally used in online systems, retrieval time is also as important as accuracy. In this paper, we propose a QBE-based MIR system and investigate the impact of automatic music genre prediction on the performance of it, specifically on perspective of accuracy-time trade-off, using a score-based genre prediction method as well as similarity measures. The proposed system is evaluated on a dataset containing 6000 music pieces from six musical genres, and we show that how much improvement on the performance can be achieved in terms of accuracy and retrieval time, compared with a typical QBE-based MIR system that uses only similarity measures to find the user-desired song.

Keywords Music information retrieval · Query-by-example · Score-based genre prediction · Similarity measure

1 Introduction

Nowadays, multimedia content is growing rapidly and numerous contents are being created in each second. Parallel to this enormous growth, advancements in digital storage technology make it possible to store thousands of documents in a small storage device. Thus, the necessity for more accurate browsing and retrieval of documents such as text, image, audio and video is unavoidable.

✉ Nastaran Borjian
nastaran.borjian@modares.ac.ir

[1] Department of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran

Actually, together with remarkable advances in image retrieval systems [1,2], the audio retrieval systems, particularly in music, have been significantly developed in the last decade [3] and many music information retrieval systems (MIR) have been proposed, e.g., [4–6]. In these systems, some musical–acoustical features are primitively extracted from a music dataset and stored. In the first step of the retrieval, the same features are extracted from a user-provided query, and then, the music dataset is searched using these features. Depending on the purpose of the retrieval, the output is presented in the form of either "some retrieved contents similar to the query" or "a retrieved target song."

Common input query types in MIR systems are example [7–10], singing [11,12] and humming [11,13]. One of the successful retrieval systems based on singing or humming is Online Music Recognition and Searching, OMRAS, which is established during the past decade [14].

Moreover, along with advances in multimedia devices in a few past decades, commercial music browsing engine has been developed rapidly, e.g., freeDB [15] and MusicBrainz [16] are two web-based query-by-example (QBE) music retrieval systems that are available online as well. The Shazam [17] and Musiwave [18] are also two well-known query-by-example music retrieval systems which have been advanced as application programming interface (API) on mobile phones and use audio fingerprinting methods to search a user-desired song playing on the radio or in the environment. Neuros is another query-by-example system which lets a user plug into an online service to search the target song by a 30-s clips from it [19].

To date, many music retrieval systems have used pitch representation or aimed to extract melody of music [20]. For example, Wei-Ho Tsai et al. [8] introduced a query-by-example system to retrieve the cover versions of songs, which searches target songs with an identical tune while

might be performed in another language or even by different signers and returns the songs having main melody similar to that of the query. Although melody extraction and pitch contour detection yield good retrieval performance in query-by-humming or query-by singing systems, usually do not perform well in query-by-example music retrieval systems, in particular, with impure queries. Thus, some other solutions are proposed to solve this problem. For instance, Hel'en and Virtanen parameterized the signal by GMM and HMM models and employ these parameters to search the database using different similarity measures in an audio query-by-example system [7]. As well, Shazam as a query-by-example system detects intensity peaks of the spectrogram to produce a spare feature set in a frequency range and aims to retrieve the queries with length of up to 15 s when even the offered music is transmitted over a mobile phone line from a noisy environment such as a nightclub. In another work, a QBE-MIR system based on sound source separation is proposed [21]. In this system, three groups of sound signals are separated from queries based on drums, guitar and vocal and then processed by volume balance control. Next, a re-mix stage that is equal to musical genre shift is performed on queries to find the retrieval results from some specific genres listed as Classical, Dance, Jazz and Rock.

However, retrieval of an audio content based on audio queries is an important and challenging issue in the research field of content-based access to popular music. Many of us have experienced to listen enthusiastically to a piece of music and to say what this sounds like, while we do not know its title or artist. To address this problem, a query-by-example music retrieval system attempts to search a popular music dataset using a fragment of desired song as input. These query-by-example music retrieval systems have been an active and attractive area of researches in the MIR, in particular, to develop the web-based music browsing engines and APIs. Therefore, terms of accuracy and retrieval time are two essential factors to yield satisfactory results in these systems.

In this paper, we propose a two-stage QBE music retrieval system using score-based music genre prediction and similarity measure, which receives a user-provided example and retrieves the target song from a musical dataset. In the first stage, a score-based method trained using features that here are mel-frequency cepstral coefficients (MFCCs) [22–24], extracted from dataset predicts the genre of the query. Then, in the second stage, the music pieces belonging into a specific genre yielded from previous stage are searched by a proposed method based on KullBack–Leibler (KL) divergence as similarity measure [7,25] to find the target song (see Fig. 1).

This work is on investigating about the effect of genre concept and we experimentally manifest how far the performance of a QBE-based MIR system can be enhanced in the accuracy and retrieval time terms by means of the automatic genre prediction, compared with a typical and musical QBE system developed only using similarity measures. Indeed, there exists an accuracy-time trade-off in the performance of a QBE-MIR system, so that accuracy may be increased using a high computational method while retrieval time is not suitable for an online system. Obviously, if genre of query is known previously, the search space on dataset and thus retrieval time is reduced significantly, but we have a trade-off on accuracy because of irrelevant genre error if user does not know the genre of query and it is necessary to predict it automatically by system. In other words, as we would expect, having perfect genre helps a lot and gives major benefits to reduce retrieval time, but in conditions that an almost perfect genre predictor is used to avoid from propagation of genre error to similarity measure stage and to obtain competitive results and thus to overcome the accuracy-time trade-off.

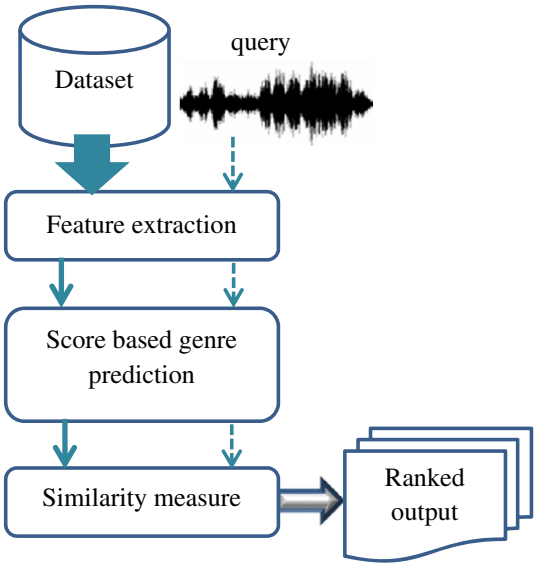

Fig. 1 Block diagram of the proposed query-by-example music retrieval system, which includes score-based genre prediction and similarity measure stages

Although numerous genre classification algorithms, for example [23,26,27], are already proposed, but we emphasize that our technique is not aimed to be a state-of-the-art on automatic musical genre recognition or genre classification system. In fact, the contribution of this study in music information retrieval field is to show the impact of "genre" as a general concept in Western music on the performance of a QBE-based MIR.

This paper is ordered as follows: The proposed QBE system is described in Sect. 2. The experimental results are presented in Sect. 3 and discussed in Sect. 4. Finally, Sect. 5 gives the conclusions.

Table 1 Dataset, music pieces and query group

Genre	Dataset	Music pieces	Query group
Classical	320	2040	1020
Electronic	115	1220	610
Jazz-Blues	26	220	110
Metal-Punk	45	400	200
Rock-Pop	101	720	360
World	122	1400	700
Total	729	6000	3000

2 The proposed QBE system

Music information retrieval (MIR) can be done using different types of queries. The most used ones are query by example, singing and humming. This paper focuses on just query-by-example (QBE) type, which gets the input as a fragment of the desired song.

The main goal of this paper is to explore the impact of automatic genre prediction in a QBE-based MIR system. Our proposed system has two stages: score-based genre prediction and similarity measure, shown in Fig. 1. The genre prediction stage predicts the genre of the query, as will be explained in Sect. 2.2, reduces the search space and consequently reduces the retrieval time. Then, the similarity measure stage searches the music pieces belonging into that specific genre and retrieves the target song as output; details will be described in Sect. 2.3.

We measure retrieval accuracy as well as retrieval time to show how much genre concept prediction improves a query-by-example-based music retrieval system in those terms.

In this work, we use the well-known Magnatune dataset [28], which contains 729 songs from six musical genres, as presented in Table 1. We constructed 6000 music pieces or clips from that dataset (see Table 1), such as every piece has a length of 30 s sampled in 44,100 Hz, and also music pieces from the same song has no overlap.

A query is defined as a random 5-s fragment of any 30-s music piece that one is also selected randomly per genre. A random query group including 3000 ones is formed to evaluate the system. Since queries are accidentally selected from each genre, thus it is possible to be multiple queries from the same music piece per genre. In other words, any substitution is allowed.

2.1 Feature extraction

So far, different representations and feature sets were proposed and tested in the music retrieval or music classification areas. An efficient representation maps the most relevant information of a music signal into a feature space depending on the performance of the system. A representation normally consists of a number of features extracted from short-time frames, typically 20–60 ms.

Since the auditory system of human has been found to perform frequency analysis [29], usually the spectrum analysis of the sound signal in comparison with the time analysis correlates well with the perception which a human has from the sound. Therefore, the representation that parameterizes the spectrum of a sound signal specifically music, has been found to achieve better results.

The most common representation associated to the genre concept is timbral texture representation consisted of temporal features, spectral features and mel-frequency cepstral coefficients (MFCCs) [22–24]. Some applications also use this representation combined to features of musical characteristics such as harmony, pitch and duration pairs, rhythm and beat [22,26]. While many current algorithms use above features, there is no clear evidence that which features are optimal to be linked to the genre concept.

Additionally, it is often observed that humans do accurately the genre prediction. This can motivate the researchers for incorporation of physiological models of human into systems developed on genre concept. Termens et al. [30] evaluated how human predicts the genre of a music signal collapsed by either timbral or rhythmical distortion. The results of this research states that human can better predict the genre of a music collapsed by timbral distortion than rhythmical one, which practically leads to a conclusion that timbral texture representation is more close to the physiological models of human auditory system. Also, George Tzanetakis et al. in [22] experienced some feature sets and showed that better results in genre classification will be obtained by the timbral texture features than musical ones such as pitch and beat. Although we recall that our aim is not genre classification and have no claim in this field, but we only follow the genre prediction of the query by a score-based method to reduce the search space. Thus, in this work we use timbral texture features and calculate 20 MFCCs [31], for every 40 ms frame while those frames have an overlap ratio of 50% (as done by a number of literature works).

2.2 Genre prediction

It is often observed that human can accurately percept the difference between music genres and identify them even when he/she does not have any musical intuition about genres. Anders Meng et al. [32] conducted a listening test on two short and long databases to estimate the ability of human to classify a music based on its genre. It was observed that the average human accuracy is 95% when different sound clips of 10-s length are listened. We inspire from this characteristic in this work and show how much the performance of a query-by-example music retrieval system will be improved using

automatic music genre prediction while this aspect of music has been generally involved to develop online QBE systems. As will be explained, we employ a score-based method using a binary CART decision tree [33,34] to predict the genre of the query, which reduces the search space on a genre-labeled dataset and thus reduces the retrieval time.

2.2.1 Decision tree

To predict the genre of the query, we built a binary CART decision tree on whole the dataset followed with Gini Diversity Index (gdi) criterion [33,34].

In a binary CART decision tree, two children nodes are created from a parent node in such a way that the learning samples related to each of the children node are purer than ones from parent node, which means the variety of classes will be lower.

There are several types of split methodologies that can be considered at each step to create children nodes. We develop a methodology that each split depends on the value of only a single feature. Let us denote the features by $x_1, x_2, \ldots, x_p$ where p is number of them. For a feature x_k, which has numerical value in our work, a subset of frames of dataset can be divided, so that one of the new created subsets has $x_k < s_k$ and the other has $x_k > s_k$ that s_k is a special threshold named split value.

Let $x_k(1) < x_k(2) < \cdots < x_k(M)$ denote the sorted distinct values of x_k observed in frames belonging into a subset that is traced to be divided. If $x_k(m)$ and $x_k(m+1)$ be any two consecutive sorted distinct values of variable x_k, taken halfway between them defined as Eq. (1) for $m = 1, 2, \ldots, M-1$ can be considered as the split value s_k [33,35].

$$T = \frac{x_k(m) + x_k(m+1)}{2} \tag{1}$$

If feature x_k is a numeric variable with M distinct values, we have to compute $M-1$ split values to determine an optimal threshold. Thus, even if there are many different ordinal features, there exist only a finite number of possible split values of this form.

To choose an optimal split value in each branch node, we use Gini Diversity Index (gdi) impurity criterion [34,36] given as:

$$\mathrm{gdi}_{\mathrm{node}} = 1 - \sum_i p^2(i). \tag{2}$$

where the sum is over the classes i at the node and $p(i)$ is the observed fraction of frames with class i that reach to the node. A node with just one class, a pure node, has Gini index 0. Otherwise, the Gini index is always positive and smaller than one in an impure node. Therefore, the Gini index is a measure of node's impurity. We recall that the class in this work means "genre."

2.2.2 Score-based genre prediction

In order to predict the genre of the query, we first train a binary CART decision tree by all frames of dataset. Then, the genre of the query is predicted by a score-based method, so that the decision tree assigns a score vector to each frame of the query. It is performed by computing the posterior probability of each class i per frame of query, Xnew, shown as $p(i|\mathrm{Xnew})$. Thus, element kth of score vector for a frame of query is:

$$\mathrm{score}_k = p(k|\mathrm{Xnew}) \quad \text{for} \quad k = 1, \ldots, G \tag{3}$$

Here, G is the number of classes.

In the following, for two best scores of each frame of query, we calculate an uncertainty measure as:

$$\frac{\mathrm{score}_{\max 1}(\mathrm{Xnew}) - \mathrm{score}_{\max 2}(\mathrm{Xnew})}{\mathrm{score}_{\max 1}(\mathrm{Xnew})} \leq \frac{1'}{2} \tag{4}$$

If this uncertainty measure is true, then a weight vector w_{Xnew} is defined for frame Xnew as:

$$w_{\mathrm{Xnew}}(j) = \begin{cases} 0.5 & \text{if } j = \arg\ \mathrm{score}_{\max 1}(\mathrm{Xnew}) \\ 0.5 & \text{if } j = \arg\ \mathrm{score}_{\max 2}(\mathrm{Xnew}) \\ 0 & \text{others} \end{cases} \tag{5}$$

where $j = 1, \ldots, G$, else if uncertainty measure is false, then w_{Xnew} is:

$$w_{\mathrm{Xnew}}(j) = \begin{cases} 1 & \text{if } j = \arg\ \mathrm{score}_{\max 1}(\mathrm{Xnew}) \\ 0 & \text{others} \end{cases} \tag{6}$$

Finally, weight vectors of all frames of the query are summed and the genre predicted for the query, y, is given by:

$$y = \arg\max \sum_{\mathrm{query}} w_{\mathrm{Xnew}} \tag{7}$$

Table 2 provides one example in details on how to map a score vector to a typical frame of the query by decision tree, if query is assumed to be selected from genre 2. If decision tree is able to assign the genre of frame truly, then score vector will be vector A, else vector B.

Uncertainty measure along with weight vector that we applied in this stage increases in overall the chance of true genre to be assigned to the query, and also decreases the chance of other genres.

Table 2 One example of score vector of a typical frame of the query

	Genre 1	Genre 2	Genre 3	Genre 4	Genre 5	Genre 6
vector A	0	0.8	0	0.1	0.1	0
vector B	0.1	0.3	0.4	0.1	0	0.1

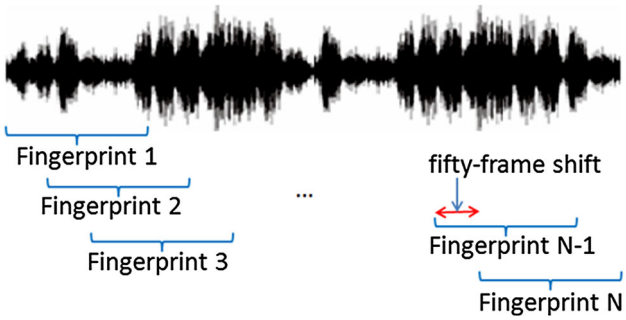

Fig. 2 Query-length fingerprints of a typical music piece

2.3 Similarity measure

After that the genre of the query is predicted, further search is done on the music pieces in the dataset having the same genre with the predicted genre. To accomplish this, the music pieces are first segmented into query-length intervals, known as fingerprint, used also, for example, by Shazam [17] and Musiwave [18], and then, the distance between each fingerprint and query is calculated by a similarity measure. The fingerprints having lower distance values are obviously more matched with the query and accordingly ones with higher distance values have less matching. The segmentation is experimentally done with fifty-frame shifts, shown in Fig. 2.

2.3.1 Distance measure

Let us denote the feature sequence matrices of fingerprint A and query Q by $A = [a_1, \ldots, a_{FA}]^{\mathrm{T}}$ and $Q = [q_1, \ldots, q_{FQ}]^{\mathrm{T}}$, respectively, where F_A and F_Q are the number of frames in each of them. To find the best-matched music pieces, the query is first moved over each music pieces, fingerprint by fingerprint, and then, Kullback–Leibler divergence [37] is calculated in each step by:

$$\mathrm{KL}\left(p_A(x)||p_Q(x)\right) = \frac{1}{2}\left[\log\frac{|\Sigma_Q|}{|\Sigma_A|} + Tr(\Sigma_Q^{-1}\Sigma_A^{)} + (\mu_A-\mu_Q)^{\mathrm{T}}\Sigma_Q^{-1}(\mu_A-\mu_Q) - N_F\right] \tag{8}$$

where μ_{A}, μ_{Q}, Σ_{A} and Σ_{Q} denote the mean and covariance matrices of feature vectors for fingerprint A and query Q, respectively. In fact, there exist an implicit assumption that the probability density function of features for each fingerprint and also query is as a multivariate Gaussian distribution, i.e., $p(x) = \mathcal{N}(x; \mu, \Sigma)$, known as normal distribution too.

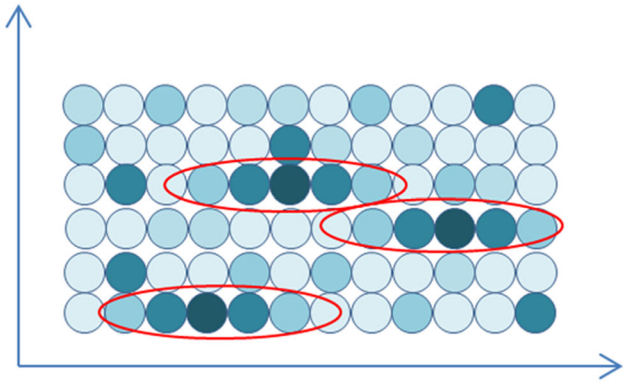

Fig. 3 Similarity matrix

After that the distance calculations are finished for each music piece in the dataset having the same genre with the query, a similarity matrix, SM, is formed as:

$$\mathrm{SM} = \begin{bmatrix} k_{1,1} & k_{1,2} & \cdots & k_{1,P} \\ k_{2,1} & k_{2,2} & \cdots & k_{2,P} \\ \vdots & & & \vdots \\ k_{S,1} & k_{S,2} & & k_{S,P} \end{bmatrix} \tag{9}$$

Here, S is the number of music pieces in the predicted genre, each which has P fingerprints. Figure 3 illustrates SM matrix in color. The darker points are fingerprints having less distance.

It is found that usually the musical format does not change suddenly during a short time interval. On the other hand, it is possible that a few seconds of a song to be repeated in another one. We consider those characteristics in SM matrix to select the best-matched music pieces and to improve the retrieval results. To perform this, in addition to three selected fingerprints having less distance, four before and four after fingerprints (see Fig. 3) are also taken and their distance values are summed with that of central fingerprint, so that a 3-element distance vector is formed and sorted in ascending order. After that, the target music piece is proposed according to index of first element of sorted distance vector. In the next sections, this technique is referred as min–sum–min manner. For more ranked output results, above algorithm is repeated, e.g., for 3-ranked results known as Top-3, firstly nine fingerprints are selected and then ranked results are formed with the corresponding music pieces whose indices are determined according to sorted distance vector.

We evaluated the performance of our QBE-based MIR system using the retrieval accuracy, defined as "the number of queries whose target songs are retrieved in the ranked outputs" divided by "the number of queries." Top-N accuracy

Table 3 Confusion matrix for score-based genre prediction stage

Actual genres	Predicted genres					
	Classical	Electronic	Jazz-Blues	Metal-Punk	Rock-Pop	World
Classical	97.30	0.55	0.57	0.4	0.53	0.65
Electronic	0.73	97.34	0.67	0.46	0.47	0.33
Jazz-Blues	2.65	0.3	94.77	0.37	1.48	0.43
Metal-Punk	1.53	0.31	0.49	96.31	0.4	0.96
Rock-Pop	0.78	0.4	0.58	0.59	97.15	0.5
World	0.84	0.54	0.69	0.42	0.53	96.98

means percentage of queries whose target songs are found to be among *N*-ranked output candidates.

3 Experimental results

In this work, we proposed a query-by-example music retrieval system through score-based genre prediction and similarity measure. As previously presented, the genre prediction here is only aimed to reduce the search space and differs from deeply meaningful "genre classification" area.

Our experiments are conducted using 6000 music pieces of 30-s length in six different genres, as presented in Table 1. Each query is a 5-s fragment selected arbitrarily from original 30-s music pieces while any repeat is also allowed to select random queries. The experiments are done using 3000 queries.

We follow three experiments to test our QBE-based MIR system. First, the score-based genre prediction stage is tested by itself. Performance of the proposed system is evaluated using accuracy and retrieval time in second experiment. Third experiment compares the performance of our system with the case that genre prediction stage is ignored.

3.1 Experiment I

The score-based genre prediction stage explained in Sect. 2.2.2 is tested here, and results are presented by "genre prediction rate," means how large percentage of the genre corresponding to the queries was correctly predicted. We achieved 96.64% genre prediction rate using decision tree together with score-based uncertainty measure.

Table 3 reports confusion matrix for score-based genre prediction stage. Each value on the matrix is the percentage of a certain genre predicted for queries (built-in columns) when queries are actually known to have a particular genre (built-in rows).

The values in the confusion matrix show that the highest confusion was clearly between the Jazz-Blues and Classical genres (2.65%), which is possibly, because they share musical elements. The Electronic and Jazz-Blues genres were found to have the best and the worst prediction rates, respectively, and the majority of the genres have been slightly confused with the Classical genre, which may be, because the classical genre has a wide variety in form and texture.

It should be noticed that the decision tree in score-based genre prediction stage is trained on whole music pieces of the dataset, thus is differed from a typical genre classification algorithm (for example, see [22,24]), where dataset is divided into two separate train and test groups and results are presented by *k*-fold cross validation.

3.2 Experiment II

Top-1 to Top-6 retrieval accuracy of the proposed QBE system is illustrated in Fig. 4. For more output candidates (N), retrieval accuracy obviously increases, as is 94.23% for Top-1 and 95.47% for Top-4, which is because the target music piece has higher chance to be one of the output candidates.

Furthermore, the retrieval accuracy has been observed to be limited to 96.64% where *N* is six, presented as Top-6, due to genre error propagation. In other words, since the proposed QBE system is sequential, the score-based genre prediction stage limits system to achieve the retrieval accuracy more than genre prediction rate, and thus, for Top-*N* candidates bigger than six, our system provides the retrieval accuracy of 96.64%, which means that the whole of the retrieval error is made by the genre prediction stage.

Figure 5 shows that how much of retrieval error is due to each of irrelevant genre and irrelevant music piece. Irrelevant music piece error means that the genre of the query and thus the group whose music pieces should be searched was correctly predicted, but the target music piece corresponding to the query has not been found and retrieved by system. For example, our system provides a retrieval accuracy of 95.47% for $N = 4$, which means from 4.53% retrieval error, the genre prediction and the similarity measure stages make 3.36 and 1.17% error rates, respectively (see Fig. 5). Since the genre prediction rate is 96.64%, thus 3.36% irrelevant genre error propagates into similarity measure stage and influences the final retrieval accuracy.

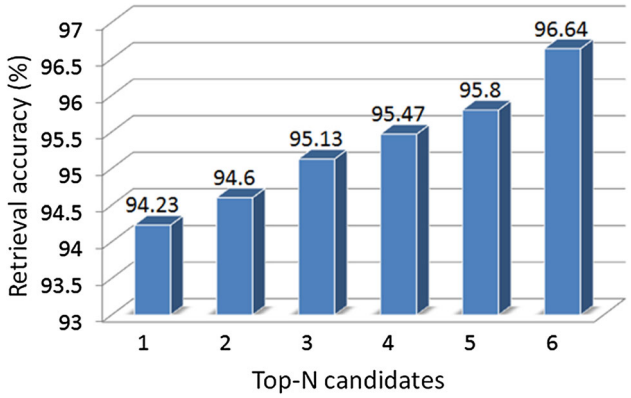

Fig. 4 Retrieval accuracy of the proposed QBE system

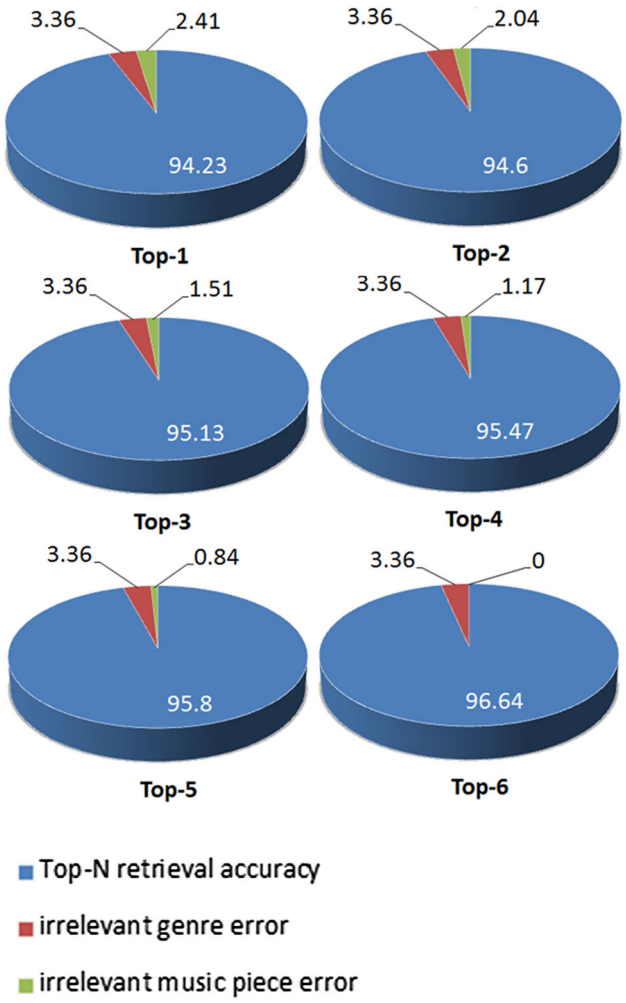

Fig. 5 Pie graph of retrieval accuracy and errors in the proposed QBE system

Retrieval time of our QBE system for all cases in Fig. 4 is almost 410 ± 170 ms, depending on the predicted genre of the query, e.g., a query predicted to be from Jazz-Blues genre has the lowest retrieval time, compared with a query from Classical genre that causes the highest one. The genre prediction stage takes 110 ms with standard deviation of ± 20 ms made by multilayer structure of the decision tree.

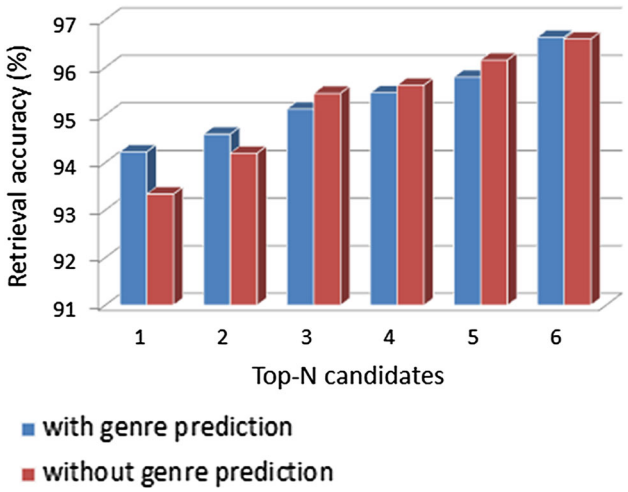

Fig. 6 Impact of genre prediction stage for proposed QBE system

The algorithms were implemented by MATLAB, and the simulations were performed using an Intel i7-3770 K–3.9 GHz Turbo Boost PC. Although, offline training of decision tree is very time consuming, but it is required to be performed once and after that, only the time for loading and running is contributed for the retrieval time.

3.3 Experiment III

To show the impact of the music genre prediction, the performance of our QBE-based MIR system which consists of a genre prediction stage along with a KL-divergence-based min–sum–min similarity measure will be compared with a case that just similarity measure is used across the whole dataset fingerprint by fingerprint to find the target music piece, i.e., the genre prediction stage is omitted. Figure 6 shows the results of this comparison for Top-1 to Top-6.

As is shown in Fig. 6, the retrieval accuracy without any genre prediction is completely competitive with that from proposed QBE-MIR using score-based genre prediction, which limits the search into a specific genre, while the retrieval time without genre prediction increases to 1730 ms, i.e., roughly in order of 4:1.

3.4 Comparisons

A perspective on state-of-the-art research on query-by-example music retrieval systems show that similarity measure stage in this work is a representative of some systems reported in the literature, for example [7,25].

3.4.1 Hel'en et al. method

As has been reported in the literature, the work done by Hel'en and Virtanen evaluated a number of distance measures for an audio query-by-example system [7], whose dataset contained different types of audio including environmental, music, sing and speech. They followed a k-nearest neighbor method using different distance measures to retrieve 10-s excerpts from the same category with the query.

Since the audio query-by-example system described in [7] differs from our work in dataset, algorithm and the aim of retrieval, real comparison is not feasible and just some of the distance measures explained in that work [7] are used here to test our QBE-based MIR system. To perform this, all music pieces on dataset are searched, fingerprint by fingerprint, and distance between each fingerprint and query is calculated in order to find the closest matches. Those music pieces having the smallest distances in one of the accompanying fingerprints are selected and after using the min–sum–min method (proposed in this work, see Sect. 2.3.1) would be retrieved and ranked while only one of them might be the target music piece. The distance measures examined here are listed as:

(i) Euclidean distance over histogram matrices constructed in eight quantization levels [38],

(ii) Mahalanobis distance defined as:

$$\text{Mahalanobis Dis.}(A, Q) = (\mu_A - \mu_Q)^{\mathrm{T}} \Sigma^{-1} (\mu_A - \mu_Q) \tag{10}$$

for which probability density function (PDF) of features for query and each fingerprint is assumed to be a normal distribution [7], and Σ denotes the covariance matrix of the features across all music pieces.

(iii) closed-form $\mathcal{L}2$-norm distance for two Gaussian mixture models (GMM) proposed in [39], when probability density function of features for each fingerprint and also query is modeled with a 12-GMM, i.e., $p(x) = \Sigma_{\mathrm{i}=1:12} w_{\mathrm{i}} \mathcal{N}(x;\ \mu_{\mathrm{i}},\ \Sigma_{\mathrm{i}})$, given by:

$$e = e_{AA} + e_{QQ} + 2e_{AQ} \tag{11}$$

$$\begin{aligned} e_{AA} &= \sum_{i=1}^{I_A} \sum_{j=1}^{I_A} w_i^A w_j^A D_{i,j,A,A} \\ e_{QQ} &= \sum_{i=1}^{I_Q} \sum_{j=1}^{I_Q} w_i^Q w_j^Q D_{i,j,Q,Q} \\ e_{AQ} &= \sum_{i=1}^{I_A} \sum_{j=1}^{I_Q} w_i^A w_j^Q D_{i,j,A,Q} \end{aligned} \tag{12}$$

where,

$$D_{i,j,k,m} = \int_{-\infty}^{\infty} p_k(x)_i \; p_m(x)_j \, \mathrm{d}x \tag{13}$$

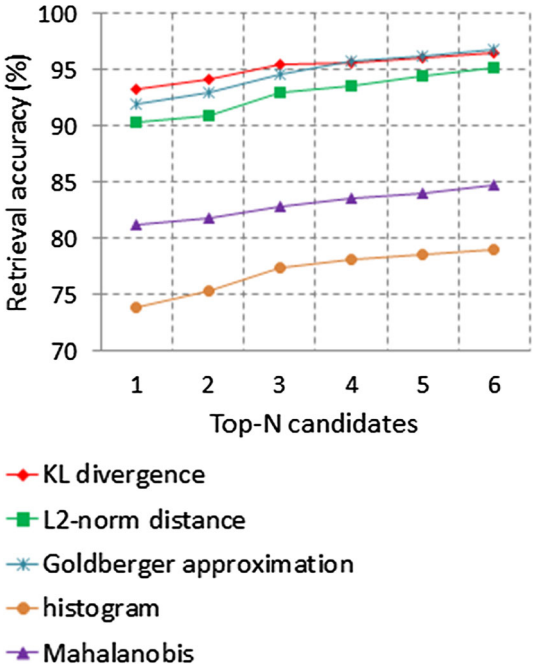

Fig. 7 Retrieval accuracy using different distance measures [7], compared with the KL-divergence-based min–sum–min similarity measure (ours)

For I_A, $I_Q = 12$ and $k, m \in \{A, Q\}$.

and (iv) Goldberger approximation of Kullback–Leibler divergence for two 12-GMM [37], given as:

$$\mathrm{KL}_{\mathrm{Goldberger}}(p_A(x)||p_Q(x)) = \sum_{i=1}^{I_A} w_i^A \left(KL(p_A(x)_i||p_Q(x)_{m(i)}) + \log \frac{w_i^A}{w_{m(i)}^Q} \right) \tag{14}$$

for,

$$m(i) = \arg\min_j KL(p_A(x)_i||p_Q(x)_j) - \log\left(w_j^Q\right) \tag{15}$$

Figure 7 illustrates the retrieval accuracy achieved by those distance measures for Top-N candidates ranged from 1 to 6, compared with the KL-divergence-based min–sum–min similarity measure described in Sect. 2.3.1.

Looking at Fig. 7, both KL divergence and Goldberger approximation give the most accurate results for all Top-N candidates ranged from 1 to 6, which shows not only KL divergence but also its approximation could model the similarities and differences in feature space well, so that those will be almost converged where N goes up to four. In comparison with the Goldberger approximation, the retrieval accuracy of $\mathcal{L}2$-norm distance is competitive and a little lower, while both have high computational cost to be used in an online music retrieval system.

Histogram method and Mahalanobis distance are observed to yield lower retrieval accuracy than others, which is possibly because both of them use information which is not enough rich to model differences. However, the retrieval accuracy achieved using statistics-based Mahalanobis distance is higher than that of histogram method, which is a numerical measure.

Although, the KL divergence, which achieves the highest accuracy in this study, assumes simply a normal distribution for each fingerprint (see Sect. 2.3.1), but our experimental results show that statistical information of fingerprint including mean and covariance is enough suitable to map similarities. Thus, the proposed KL-divergence-based min–sum–min similarity measure has been found to perform clearly better than others do, whereas it imposes a low complexity cost to online systems.

An important problem with the distribution-based methods is that we assume the features are independent and a full-covariance matrix can be easily calculated for them. While this assumption is not necessarily true, especially for a 5-s short signal, included 250 frames in our work, so that the features are slightly dependent to each other and thus a singular covariance matrix for some of the fingerprints or even queries is unavoidable. The Metal-Punk genre has been observed to reveal this problem more than other ones. To address this problem, it is possible to assign a predefined value to the fingerprint's determinant if it is smaller than a threshold. These parameters are better to be adapted for each genre. Another solution, which applied in this work, is to discard the fingerprints with a singular covariance matrix. Since the step-size between fingerprints is very low, 50 frames in our simulations, eliminating a number of fingerprints did not make a significant effect on the final retrieval results.

Two main questions about an online music retrieval system are that how much accuracy is necessary for such as system and also how much retrieval time is acceptable, while there exist no clear evidence to answer those questions. It may be because of variety in dataset, search algorithm and output results, whereas the robustness against noise and distortions should be considered.

For example, a comparison between Shazam and SoundHound [40], two commercial mobile QBE music retrieval services, is done and shown the Shazam was faster than SoundHound, while the accuracy of SoundHound is more than Shazam. The SoundHound was accurate enough and had a 95+ % success ratio with 21 s of time for preparing the answer. The Shazam shown an accuracy of almost 85% with 12 s for discovering time [41]. In overall, it seems that an accuracy of 95% is adequate to acquire a competitive marketing research for query-by-example music retrieval systems, although an error rate of 5% is still too high for verification such as services [42].

Table 4 Retrieval time for different distance measures [7], compared with the KL-divergence-based min–sum–min similarity measure (ours)

Distance measure	Retrieval time (ms)
KL divergence	1730
$\mathcal{L}2$-norm distance	2850
Goldberger approximation	2730
Histogram method	1340
Mahalanobis distance	1680

Table 4 includes the retrieval time for different distance measures, examined according to Fig. 7.

The best retrieval time is obtained with histogram method and next that Mahalanobis distance, while $\mathcal{L}2$-norm distance has the most computational cost. However, in addition to retrieval accuracy, the retrieval time is an essential parameter to develop an online and popular music information retrieval system.

Obviously, retrieval time is dependent upon dataset size and this dependency is approximately linear. We remind that retrieval times mentioned in Table 4 are for a dataset containing 6000 clips, each has 30 s long (see Table 1). Thus, it is easy to linearly scale the retrieval time corresponding to each method in Table 4 with either increasing dataset size or when length of clips is different.

To summary, the performance of a typical QBE-based MIR system is highly dependent on the distance measure, so that the histogram method exhibits the results obviously worse than ones obtained with others while its computational cost is low. On the other hand, Goldberger approximation performs well, but is not very practical in this work due to high complexity cost.

In addition, a specific distance measure is likely to perform better over a particular genre, thus we organized to evaluate each genre separately by the different distance measures in our QBE system, but the experimental results showed that the proposed KL-divergence-based min–sum–min still outperforms over all genres.

3.4.2 Harb et al. method

As well, we develop the KL-divergence-method applied in [25] to compare that in retrieval accuracy term with min–sum–min manner proposed in this study. The search here is also done fingerprint by fingerprint, and we implement just the global similarity defined for that work, which expresses the similarity between songs as from the same musical genre done by a k-NN classification manner for Top-30 [25]. Therefore, a reasonable comparison between the results of that work and ones from our approach is impossible because of lack of common dataset and meaningfully the aim of systems.

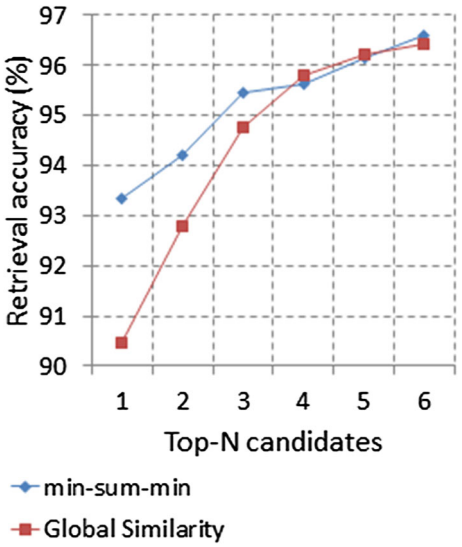

Fig. 8 Retrieval accuracy using min–sum–min manner (ours) and global similarity [25]

The work explained in [25] uses distances between statistical distributions of the spectral features extracted by means of the short-term Fourier transform and then filtered by a filter bank containing 20 filters distributed based on the Mel scale, and defines the global similarity as the average of MD, MDHF, MDLF and SD, which are:

- Min Distance (MD): the average of the three min values of the SM,
- Min Distance for High Frequencies (MDHF): the MD of the SM for frequencies between 1 and 4 kHz,
- Min Distance for Low Frequencies (MDLF): the MD of the SM for frequencies lower than 1 kHz,
- Sum Distance (SD): the average of all values of the SM.

Figure 8 illustrates the comparison between min–sum–min manner (proposed in this work) and global similarity. With respect to Fig. 8, the min–sum–min manner yields the retrieval accuracy as good as one from global similarity in lower computational cost. This is possibly, because we used MFCCs that implicitly reflect the frequency information rather than filtered spectral features. On the other hand, instead of accounting three min points of SM alone, the local correlation of fingerprints placed adjacent to them is also considered.

4 Discussion

Our contribution to the music information retrieval is to evaluate the impact of automatic genre prediction of the query on performance of a QBE-based MIR system in accuracy and retrieval time terms. To accomplish this, some different distance measures are evaluated in a fingerprint by fingerprint manner and their results are compared with those from our proposed system which limits search into a predicted genre using a score-based genre prediction algorithm and also applies a KL-divergence-based min–sum–min similarity measure to find the target music piece.

The experimental results reveal although the final retrieval accuracy of the proposed system is competitive in comparison with a typical online QBE-based MIR system given in literature and there seems to be no significant improvement in accuracy term, but it should be noticed that this achievement is attained while time consumption is reduced roughly in order of 1/4. Of course, this is partly due to the good performance of score-based genre prediction. As mentioned previously, here the whole dataset is used as training set and thus our application is completely differed from "genre classification" field.

As well, the further optimization of the proposed system, for example using feature and decision fusion techniques, likely improves the output performance in both terms of speed and accuracy. On the other hand, although the offline trained decision tree used for genre prediction can speed up the music retrieval but suffer from retraining, if the dataset is intended to be extended.

In our system, the error rate of the genre prediction is propagated to the similarity measure stage and has a direct effect on the final accuracy. As mentioned before, the feature and decision fusion techniques can be employed to solve this problem and optimize the manner that the genre of the query is predicted.

Generally, the proposed approach in this work assumes that the music collection/dataset is already organized by genre. Although it might not be a valid assumption in reality, but the "genre concept" is too popular to indicate and classify the variety of Western music. However, it is possible to use a large-scale dataset instead of 6000 music piece dataset, if the songs are previously classified into different genres.

Obviously, depending on the research area, features have a significant impact. MFCCs mostly reflect the spectral characteristics of the music signals, especially energy of the fundamental frequencies, which the instruments playing the music in a particular genre naturally affect those. In addition, although the decision tree trained by frame-wise MFCCs has been found to be good as a genre predictor in this work, it likely will be restricted for other generic music retrieval methods, e.g., ability to find the target song based on subjective topic, artist or even excited emotion.

We also tested a case that in each frame, MFCCs were concatenated with some temporal and spectral features summarized as spectral centroid, spectral flux, spectral roll-off, zero-crossing, minimum and maximum amplitude, first and second peaks of local energy and total energy to build a 29-

dimension feature vector per frame. In comparison with our system, the performance in this case was very competitive while the retrieval operation took a longer time. In fact, no significant improvement in this case may show that there exists redundancy in feature extraction.

In this work, query is assumed to be a continuous 5-s fragment of the target music piece and thus any alignment is not necessary to be done, while different distortions may occur for query in reality which is better to be considered in future works.

The retrieval time in this system is mostly spent in the similarity measure stage and for a large dataset with too many songs belonging to a particular genre, it becomes uncontrollable. One solution is to handle a suitable search manner with respect to the song rather than a fingerprint-wise search. Use of clustering algorithms or supervised classifiers for songs in a specific genre prior to the search will also help to solve this problem.

5 Conclusions

In this paper, a query-by-example music retrieval system was proposed which applied a score-based algorithm in system back end to predict the genre of the query and then restricted search of the corresponding target music only into the predicted genre of the dataset by a KL-divergence-based min–sum–min similarity measure, thus significantly reduced the retrieval time. The performance of such as systems that usually are used online is evaluated with accuracy as well as the retrieval time and experimental results showed that the proposed approach can achieve results with comparable accuracy in reduced query response times. Thus, it can be concluded that the automatic genre prediction can be a good solution to organize an accurate and fast QBE-based MIR system, but since the performance of that is limited by error in genre prediction stage; thus, further optimization should be done to avoid from genre error propagation.

References

1. Dharani T, Aroquiaraj IL (2013) A survey on content based image retrieval. In: International conference on pattern recognition, informatics and mobile engineering (PRIME 2013) Salem, USA, pp 485–490
2. Gong Y, Lazebnik S, Gordo A, Perronnin F (2013) Iterative quantization: a procrustean approach to learning binary codes for large-scale image retrieval. IEEE Trans Pattern Anal Mach Intell 35(12):2916–2929
3. Downie JS. The International Society of Music Information Retrieval. http://www.ismir.net/
4. Ras ZW, Wieczorkowska A (2010) Advances in music information retrieval, 1st edn. Springer, Berlin
5. Schedl M, Gómez E, Urbano J (2014) Music information retrieval: recent developments and applications. Found Trends Inf Retr 8(3):127–261
6. Casey MA, Veltkamp R, Goto M, Leman M, Rhodes C, Slaney M (2008) Content-based music information retrieval: current directions and future challenges. Proc IEEE 96(4):668–696
7. Helén M, Virtanen T (2010) Audio query by example using similarity measures between probability density functions of features. EURASIP J Audio Speech Music Process, pp 1–12
8. Tsai W-H, Yu H-M, Wang H-M (2005) Query-by-example technique for retrieving cover versions of popular songs with similar melodies. In: 6th international conference on music information retrieval, London, UK. September 11–15, pp 183-190
9. Suyoto ISH, Uitdenbogerd AL, Scholer F (2007) Effective retrieval of polyphonic audio with polyphonic symbolic queries. In: MIR '07 Proceedings of the international workshop on multimedia information retrieval, pp 105–114
10. Makhoul J, Kubala F, Leek T, Liu D, Nguyen L, Schwartz R, Srivastava A (2000) Speech and language technologies for audio indexing and retrieval. Proc IEEE 88(8):1338–1353
11. Tsai W-H, Tu Y-M, Ma C-H (2012) An fft-based fast melody comparison method for query-by-singing/humming systems. Pattern Recogn Lett 33:2285–2291
12. Yu H-M, Tsai W-H, Wang H-M (2008) A query-by-singing system for retrieving karaoke music. IEEE Trans Multimed 10(8):1626–1637
13. Unal E, Chew E, Georgiou PG, Narayanan SS (2008) Challenging uncertainty in query by humming systems: a fingerprinting approach. IEEE Trans Audio Speech Lang Process 16(2):359–371
14. Kaminskas M, Ricci F (2012) Contextual music information retrieval and recommendation: state of the art and challenges. Comput Sci Rev 6:89–119
15. Schröder A, Keith M. Free database. http://www.freedb.org
16. Kaye R. The Open Music Encyclopedia. https://musicbrainz.org
17. Barton C, Inghelbrecht P, Wang A, Mukherjee D. Shazam Company. http://www.shazam.com/company
18. Chuffart F. Musiwave. http://www.musiwave.net
19. Born J. Neuros. www.neurostechnology.com
20. Salamon J, Gómez E (2010) Melody extraction from polyphonic music signals using pitch contour characteristics. IEEE Trans Audio Speech Lang Process 20(6):1759–1770
21. Itoyama K, Goto M, Komatani K, Ogata T (2010) Query-by-example music information retrieval by score informed source separation and remixing technologies. EURASIP J Adv Signal Process 2010:1–14
22. Tzanetakis G, Cook P (2002) Musical genre classification of audio signals. IEEE Trans Speech Audio Process 10(5):293–302
23. Silla CN, Koerich AL, Kaestner CAA (2008) Feature selection in automatic music genre classification. In: Tenth IEEE international symposium on multimedia (ISM 2008), Berkeley, CA, pp 39–44
24. Dehkordi MB (2014) Music genre classification using spectral analysis and sparse representation of the signals. J Signal Process Syst 74:1–8
25. Harb H, Chen L (2003) A query by example music retrieval algorithm. In: Proceedings of the 4th European workshop on image analysis for multimedia interactive services (WIAMIS '03), pp 1–7
26. Chathuranga D, Jayaratne L (2012) Musical genre classification using ensemble of classifiers. In: Fourth international conference on computational intelligence, modelling and simulation (CIMSIM 2012), Kuantan, pp 237–242
27. Wang L, Huang S, Wang S, Liang J, Xu B (2008) Music genre classification based on multiple classifier fusion. In: Fourth international conference on natural computation, pp 580–583
28. ISMIR audio description contest. http://ismir2004.ismir.net/genre_contest/index.htm

29. Zwicker E, Fastl H (2013) Psychoacoustics: facts and models. Springer, Berlin
30. Termens EGi (2009) Audio content processing for automatic music genre classification: descriptors, databases, and classifiers. Ph.D. thesis, Department of Information and Communication Technologies, University Pompeu Fabra, Barcelona
31. Rabiner LR, Juang BH (1993) Fundamental of speech recognition prentice, 1st edn. Prentice Hall, Prentice
32. Meng A, Ahrendt P, Larsen J (2007) Temporal feature integration for music genre classification. IEEE Trans Audio Speech Lang Process 15(5):1654–1664
33. Porter FC, Narsky I (2013) Statistical analysis techniques in particle physics, fits, density estimation and supervised learning. Wiley, London
34. Rokach L, Maimon O (2008) Data mining with decision trees: theory and applications. World Scientific, Singapore
35. Sutton CD (2005) Classification and regression trees, bagging, and boosting. Handb Stat 24:303–329
36. Barros RC, Basgalupp MP, Carvalho ACPLF, Freitas AA (2012) A survey of evolutionary algorithms for decision-tree Induction. IEEE Trans Syst Man Cybern Part C Appl Rev 42(3):291–312
37. Goldberger J, Gordon S, Greenspan H (2003) An efficient image similarity measure based on approximations of KL divergence between two Gaussian mixtures. In: Proceedings of the 9th IEEE international conference on computer vision (ICCV '03). Nice, France, pp 487–493
38. Kashino K, Kurozumi T, Murase H (2003) A quick search method for audio and video signals based on histogram pruning. IEEE Trans Multimed 5(3):348–357
39. Helén M, Virtanen T (2007) Query by example of audio signals using Euclidean distance between Gaussian mixture models. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP '07). Honolulu, Hawaii, USA, pp 225–228
40. Mohajer K, Emami M, Hom J, McMahon K, Stonehocker T, Lucanegro C, Mohajer K, Arbabi A, Shakeri F. www.soundhound.com
41. Gowan M. http://www.techhive.com/
42. Cox I, Miller M, Bloom J, Fridrich J, Kalker T (2007) Digital watermarking and steganography, 2nd edn. Morgan Kaufmann, Los Altos