FAST MUSIC INFORMATION RETRIEVAL WITH INDIRECT MATCHING

Takahiro Hayashi[*]*, Nobuaki Ishii*[*]*, Masato Yamaguchi*[*]

[*] Department of Information Engineering,
Faculty of Engineering,
Niigata University
8050 Ikarashi-2-no-cho, Nishi-ku,
Niigata-shi, Niigata, 950-2181, Japan

ABSTRACT

This paper presents a fast content-based music information retrieval method. The high computational cost of similarity evaluation based on musical features between a pair of music clips is a crucial problem especially for searching large music database. To reduce the computational time in similarity evaluation process, the proposed method adopts an approach called indirect matching. In the approach, a small number of music clips called representative queries, which are randomly selected from a database, are used for fast computation. As an offline process, the similarities of each music clip in the database to the representative queries are recorded as a similarity table. In the online phase, the similarity between the actual query (the music clip given by a user) and each music clip in the database is quickly estimated by referring the similarity table. Experimental results have shown that the execution time of retrieval can be greatly reduced by the indirect matching without much deterioration of retrieval accuracy.

***Index Terms*—** Music information retrieval, content-based retrieval, similarity evaluation of music, indirect matching.

1. INTRODUCTION

The rapid growth of digital audio music has generated a great deal of interest in effective ways of music information retrieval. A large number of methods for music information retrieval have been proposed [1–3]. Content-based retrieval has been regarded as a promising approach for music information retrieval [4, 5]. For efficient CBMIR (Content-Based Music Information Retrieval), many descriptors for representing the characteristics of music and metrics for similarity evaluation have been proposed.

CBMIR studies mainly focus on achieving high retrieval accuracy. However, they focus less on the computational cost in the similarity evaluation process for a large music database. Generally, when richer descriptors for well capturing the characteristics of music are used, the computational cost of the similarity evaluation increases. Therefore, development of an efficient matching method is a crucial issue for practical music information retrieval with a large database.

This paper proposes a fast similarity evaluation method called *indirect matching*. The proposed method is a general framework which is independent from individual music descriptors and metrics for similarity evaluation. The focus of the proposed method is to improve the performance of an existing CBMIR system with keeping the merits of the descriptors and metric used in the system.

2. CONVENTIONAL MUSIC INFORMATION RETRIEVAL METHODS

Music information retrieval methods can be classified into two types: tag-based retrieval and content-based retrieval. Tag-based retrieval searches music clips from a database using a text query, where text-tags have been previously given manually to each music clip in the database. The effectiveness of tag-based retrieval heavily depends on the ability to specify meaningful keywords, which may not always be possible. Directly searching acoustic music content is able to effectively compensate for the lack of reliable annotations and help improve the quality of retrieving music when manually-labeled annotations are ambiguous or missing. However, although CBMIR appears to be the only solution when manually-labeled annotations are ambiguous and missing, a lot of challenges and difficulties have to be faced.

The tag-based music information retrieval systems are fast enough because various speed-up techniques developed for text-document retrieval can be easily applied to the systems. However, there is no general framework for improving retrieval speed of CBMIR systems because individual CBMIR systems have different matching mechanisms depending on their music descriptors and the metrics for similarity evaluation.

One of frequently used music descriptors are acoustic features such as spectral rolloff and flux [4, 6, 7]. Generally, for capturing the short-time change of a non-stationary audio signal, it is divided into multiple segments in advance and acous-

tic features are extracted from each segment. Since a feature vector calculated to each segment can be mapped onto a point in the feature space, one audio signal is represented as a set of points in the feature space like $\{\mathbf{f}(1), \mathbf{f}(2), \cdots \mathbf{f}(n)\}$, where $\mathbf{f}(i)$ is the feature vector of the i-th segment of the signal.

The similarity between a pair of music clips is evaluated by matching the two distributions of points. As a metric for evaluating the similarity between two distributions, EMD (Earth Mover's Distance) is often used [4, 9, 10]. EMD can evaluate the minimum cost for transporting some commodity from supply points to consumption points [8]. In music retrieval, the minimum transportation cost can be interpreted as the minimal deformation cost from one distribution into the other.

The EMD between two music clips p_1 and p_2 is defined as follows:

$$\mathrm{EMD}(p_1, p_2) = \max(\mathrm{emd}(p_1, p_2), \mathrm{emd}(p_2, p_1)), \quad (1)$$

$$\mathrm{emd}(p_1, p_2) = \frac{\min\left(\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} v_{ij} c_{ij}\right)}{n_1}, \quad (2)$$

subject to

$$\sum_{j=1}^{n_2} v_{ij} \leq 1, \quad (3)$$

$$\sum_{i=1}^{n_1} v_{ij} \leq 1, \quad (4)$$

$$\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} v_{ij} = n_1 \geq 0, \quad (5)$$

where n_1 and n_2 are the numbers of segments in music clips p_1 and p_2, respectively. v_{ij} is the total amount of weight moved from feature point $\mathbf{f}_1(i)$ in music clip p_1 to feature point $\mathbf{f}_2(j)$ in music clip p_2. Here, each point in p_1 and p_2 initially has the weight value 1. c_{ij} represents the unit transportation cost from point $\mathbf{f}_1(i)$ to point $\mathbf{f}_2(j)$ defined as follows:

$$c_{ij} = c(\mathbf{f}_1(i), \mathbf{f}_2(j)) = \|\mathbf{f}_1(i) - \mathbf{f}_2(j)\|_1. \quad (6)$$

The notation of $\|\cdot\|_1$ means calculating L_1 norm.

While matching of distributions of feature vectors with EMD is efficient to achieve high retrieval accuracy, the computational cost can be enormous for a practical music retrieval system with a large database.

3. INDIRECT MATCHING

To reduce the computational cost of similarity evaluation, this paper proposes an efficient matching method called *indirect matching*. The approach can avoid the high-cost computation of the direct EMD calculations between a query and each music clip in a database using previously evaluated similarity results to a small number of pre-selected music queries called *representative queries*.

Before the actual online retrieval phase starts, using pre-selected representative queries $\mathbf{R} = \{r_1, r_2, \cdots, r_M\}$, the proposed method calculates in advance the similarities of each music clip in the database to the representative queries. The calculated similarities are recorded as a *similarity table* represented as follows:

$$T = [\mathbf{v}_1(\mathbf{R}), \mathbf{v}_2(\mathbf{R}), \cdots, \mathbf{v}_N(\mathbf{R})], \quad (7)$$

where N is the number of the music clips in the database and $\mathbf{v}_i(\mathbf{R})$ is defined as

$$\mathbf{v}_i(\mathbf{R}) = [\mathrm{EMD}(p_i, r_1), \mathrm{EMD}(p_i, r_2), \cdots, \mathrm{EMD}(p_i, r_M)]^T. \quad (8)$$

The i-th row of the similarity table (i.e., $(\mathbf{v}_i(\mathbf{R}))$ is called *the similarity vector* of music clip p_i to representative queries $\mathbf{R}$.

In the online phase, firstly, the similarity vector between the actual query (the user-inputted query) and the representative queries is calculated. The similarity vector of query q to representative queries $\mathbf{R}$ can be represented as follows:

$$\mathbf{v}_0(\mathbf{R}) = [\mathrm{EMD}(q, r_1), \mathrm{EMD}(q, r_2), \cdots, \mathrm{EMD}(q, r_M)]^T. \quad (9)$$

Next, the similarity between the query and each music clip in the database is indirectly evaluated by measuring the similarity of similarity vectors between the query and each music clip in the database with L_1 norm as follows:

$$i\mathrm{EMD}(q, p_i|\mathbf{R}) \equiv \|\mathbf{v}_0(\mathbf{R}) - \mathbf{v}_i(\mathbf{R})\|_1. \quad (10)$$

The symbol 'i' in 'iEMD' come from the word 'indirect'.

The proposed method assumes that if query q and music clip p_i in the database are similar (i.e. $\mathrm{EMD}(q, p_i) \simeq 0$), to another music clip r_j, $\mathrm{EMD}(q, r_j)$ and $\mathrm{EMD}(p_i, r_j)$ have similar values. That is,

$$i\mathrm{EMD}(q, p_i|r_j) = |\mathrm{EMD}(q, r_j) - \mathrm{EMD}(p_i, r_j)| \simeq 0 \quad (11)$$

is satisfied. We call $i\mathrm{EMD}(q, p_i|r_j)$ the *indirect earth mover's distance* between q and p_i to r_j. The indirect earth mover's distance can be expanded to a set of representative queries as shown in (10). That is, $i\mathrm{EMD}(q, p_i|R) \simeq 0$ is also satisfied if q and p_i are similar. On the other hand, if q and p_i are not similar, $i\mathrm{EMD}(q, p_i|R)$ becomes larger as well. Since $\mathrm{EMD}(q, p_i)$ and $i\mathrm{EMD}(q, p_i|R)$ has correlation, $i\mathrm{EMD}(q, p_i|R)$ can be substituted for $\mathrm{EMD}(q, p_i)$.

For calculating $i\mathrm{EMD}(q, p_i|R)$, the information of $\mathbf{v}_0(\mathbf{R})$ and $\mathbf{v}_i(\mathbf{R})$ is needed. Since $\mathbf{v}_i(\mathbf{R})$ can be obtained by referring the i-th row of the similarity table, only the calculation for $\mathbf{v}_0(\mathbf{R})$ is needed in run time. The computational cost of the calculations for obtaining the information of $\mathbf{v}_0(\mathbf{R})$ is greatly smaller than the direct calculations of EMD between the query and each music clip in the database because

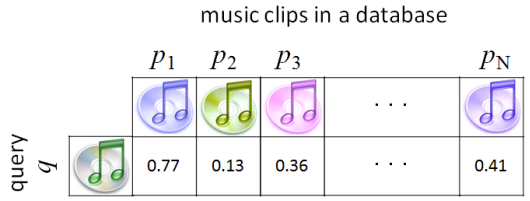

Fig. 1. Direct matching (conventional method)

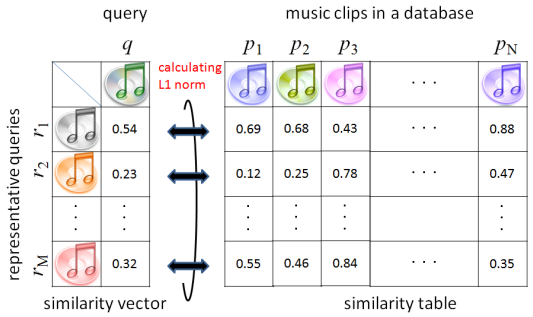

Fig. 2. Indirect matching (proposed method)

$M \ll N$. Therefore, the execution time of the proposed system can be greatly reduced.

Fig.1 and Fig.2 give intuitive explanations of the direct matching (a conventional retrieval method) and the indirect matching (the proposed retrieval method). As shown in Fig.1, the direct matching directly evaluates the EMD between the given query and each music clip in the database. The EMD calculations are repeated N times. On the other hand, the indirect matching avoids the direct EMD calculation between the query and each music clip in the database. Instead of the direct EMD calculation, the similarities (the similarity vector) between the given query and each representative query are calculated in the retrieval phase, and then the similarity of similarity vectors of the query and each music clip in the database is evaluated by calculating the L_1 norm. In the indirect matching, the EMD calculations are repeated only M times. Since $M << N$, the number of EMD calculations can be greatly reduced.

4. FEATURE EXTRACTION

Acoustic features used in this study are shown in Table 1. These features were introduced by Lu [6] and Jiang [7] and have been used in various applications such as [4]. This section briefly introduces these features.

For extracting these features, each music clip is monauralized and divided into nonoverlapping segments. In each segment, an octave-scale filter-bank is used to divide the spectrum into several subbands in order to get more details of

Table 1. Features used

Dim.	Symbol	Description
1-7	$f_2(t), \cdots, f_7(t)$	intensity of each subband
8	$f_8(t)$	overall intensity
9	$f_9(t)$	spectral centroid
10	$f_{10}(t)$	spectral width
11	$f_{11}(t)$	spectral rolloff
12	$f_{12}(t)$	spectral flux
13-19	$f_{13}(t), \cdots, f_{19}(t)$	spectral peak of subband
20-26	$f_{20}(t), \cdots, f_{26}(t)$	spectral valley of subband
27-33	$f_{27}(t), \cdots, f_{33}(t)$	spectral contrast of subband

spectrum as follows:

$$\left[0, \frac{F_0}{2^{n-1}}\right), \left[\frac{F_0}{2^{n-1}}, \frac{F_0}{2^{n-2}}\right), \cdots, \left[\frac{F_0}{2^2}, \frac{F_0}{2^1}\right), \tag{12}$$

where F_0 is the sampling rate and n is the number of subband filters.

Seven subbands are segmented from a power spectrum calculated by short-time FFT, $F(t, f)$, for each segment, where t and f are the segment and frequency indices, respectively.

Intensities of each subband, $f_1(t), f_2(t), \cdots, f_7(t)$, and overall intensity $f_8(t)$ are defined as follows:

$$f_i(t) = \sum_{f=F_L(i)}^{F_H(i)} F(t, f), \qquad (i = 1, 2, \cdots, 7) \tag{13}$$

$$f_8(t) = \sum_{f=1}^{\frac{F_0}{2}} F(t, f), \tag{14}$$

where $F_{L(i)}$ and $F_{H(i)}$ are the lower and upper bounds of the i-th subband, respectively.

Spectral centroid $f_3(t)$, spectral width $f_4(t)$, spectral rolloff $f_5(t)$ and spectral flux $f_6(t)$, which capture the shape of power spectrum, are defined as follows:

$$f_9(t) = \frac{\sum_{f=1}^{\frac{F_0}{2}} F(t, f) f}{f_1(t)}, \tag{15}$$

$$f_{10}(t) = \frac{\sum_{f=1}^{\frac{F_0}{2}} F(t, f)(f - f_3(t))^2}{f_1(t)}, \tag{16}$$

$$\sum_{f=1}^{f_{11}(t)} F(t, f) = 0.95 f_1(t), \tag{17}$$

$$f_{12}(t) = \sum_{f=1}^{\frac{F_0}{2}} \left(logF(t, f) - logF(t-1, f)\right)^2. \tag{18}$$

The strength of spectral peaks and spectral valleys of each subband are estimated by the average of a percent of the

largest values and the lowest values in the spectrum, respectively, instead of the exact maximum and minimum values. Suppose the power spectrum of j-th subband is

$$\{F_j(t,1), F_j(t,2), \cdots, F_j(t,N_j)\}.$$

After sorting it in a descending order, the new vector can be represented as

$$\left\{F'_j(t,1), F'_j(t,2), \cdots, F'_j(t,N_j)\right\},$$

where $F'_j(t,1) \geq F'_j(t,2) \geq \cdots \geq F'_j(t,N_j)$. Thus, the strength of the spectral peaks of each subband, $f_{13}(t)$ to $f_{19}(t)$, and spectral valleys of each subband, $f_{20}(t)$ to $f_{26}(t)$, and spectral contrasts of each subband, $f_{27}(t)$ to $f_{33}(t)$, are respectively defined as follows:

$$f_{12+j}(t) = \frac{\sum_{f'=1}^{\alpha N_j} \log F'_j(t,f')}{\alpha N_j} \quad (j = 1,2..,7), \quad (19)$$

$$f_{19+j}(t) = \frac{\sum_{f'=(1-\alpha)N_j}^{N_j} \log F'_j(t,f')}{\alpha N_j} \quad (j = 1,2..,7), \quad (20)$$

$$f_{27+j}(t) = f_{12+j} - f_{19+j} \quad (j = 1,2..,7), \quad (21)$$

where α is a parameter for extracting stable peak and valley values, which is set to 0.2 [7].

5. EXPERIMENTAL RESULTS

In order to evaluate the retrieval accuracy and speed of the indirect matching, we experimentally compared the indirect and direct matching methods. In the experiments, GTZAN genre correction dataset [11] was used. The dataset contains $1,000$ music clips classified into 10 genres (100 music clips per each genre). Each music clip is 30 seconds long and recorded as a 22050Hz Mono 16-bit audio file in .wav format.

The experiments for evaluation of retrieval accuracy were conducted in a leave-one-out manner, which used one music clip in the dataset as a query once at a time to retrieve the remaining 999 clips in the database, and then rotated through all the music clips. To a query, the music clips in the same genre of the query are defined as correct results. The retrieval accuracy was evaluated as the number of the correct music clips in the top N retrieval results.

Fig. 3 shows the retrieval accuracies of the direct matching and the indirect matching using 5, 20 and 50 representative queries. The representative queries was selected randomly from the database. As shown in the figure, the retrieval accuracies of the indirect matching are catching up the direct matching as the number of representative queries (M) increases.

However, we have confirmed that the retrieval accuracy changes depending on representative queries. Fig. 4 shows

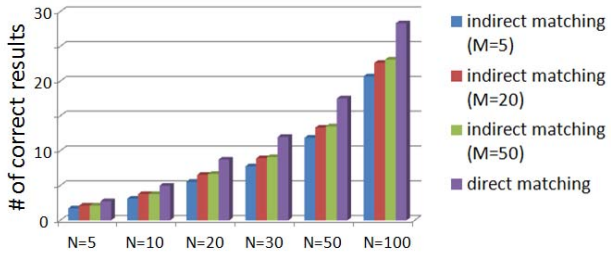

Fig. 3. Comparison of retrieval accuracies between the direct and indirect matching methods. y-axis shows the number of the correct music clips in the top N retrieval results.

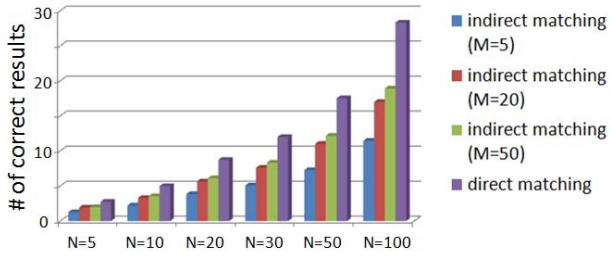

Fig. 4. The worst case of retrieval accuracy of the indirect matching when changing representative queries

the worst case of the retrieval accuracies obtained by changing representative queries. In this case, the retrieval accuracies of the indirect matching are worsen significantly from the aforementioned results in Fig. 3.

Generally, a high retrieval accuracy can be obtained when the representative queries have great variation. Therefore, for achieving high retrieval accuracy, evaluation of the variation of representative queries would be important. The techniques of the MDS (multi-dimensional scaling) would be effective for obtaining a better set of representative queries assuring the variation. Development of a selection method for good representative queries remains as a future work.

Next, we confirmed the retrieval speed of the indirect matching and its scalability to database size. To evaluate scalability, regarding the $1,000$ music clips of the GTZAN dataset as one set, we registered $1, 10, 10^2, \cdots, 10^5$ sets (i.e., from 10^3 to 10^8 music clips) to a database respectively to create various size of databases. The experiments for retrieval speed evaluation were conducted using an Intel Core i5-3210M 2.50GHz computer with 8 GB memory.

Fig. 5 shows the relation between the execution time and the database size (the number of music clips in the database).

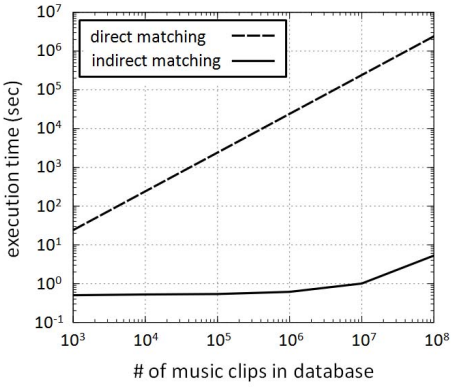

Fig. 5. Scalability evaluation of indirect matching

To a large database, the execution time of the direct matching is difficult to measure because of the extremely long execution time. Therefore, the results of the direct matching in the figure are expectation times which are defined as $T_1 + kT_2$, where T_1 is the execution time needed for feature extraction from the query, T_2 is the execution time of the direct matching when registering one music set ($1,000$ music clips) to the database, and k is the number of music sets registered in the database. As shown in the figure, even when the database size is 10^7, the indirect matching can finish retrieval within about 1 sec, which is $10^5 \sim 10^6$ faster than the direct matching.

The two experiments for retrieval accuracy evaluation and retrieval speed evaluation indicate that the execution time is greatly improved by the indirect matching without much deterioration of retrieval accuracy.

6. CONCLUSION

This paper proposes a fast music information retrieval method called indirect matching using representative queries. Experimental results have shown that the execution time is greatly improved by the indirect matching without much deterioration of retrieval accuracy.

Development of confidence measures regarding the selection of representative queries remains as a future work. These measures would be useful not only for music information retrieval but also for hierarchical audio classification. In those cases, instead of performing direct fine audio classification, rougher classification is attempted with the use of intermediate classification levels through hierarchical classification taxonomies. The common difficulty here is that how can be ensured that some queries are really representative, avoiding a potential retrieval (classification) error.

The indirect matching with representative queries can be applied to various kinds of database. We are planning to apply the proposed method to other databases as a future work.

Acknowledgements

This research was supported by JSPS KAKENHI Grant Numbers 23700109, 26330130.

REFERENCES

[1] R. Typke, F. Wiering and R.C. Veltkamp, "A survey of music information retrieval systems," *Proc. Int. Soc. Music Information Retieval Conf.*, London, UK, 2005, pp. 153-160.

[2] P. Kness and M. Schedl, "A survey of music similarity and recommendation from music context data," *ACM Trans. Multimedia Computing, Communications and Applications*, Vol. 10, No. 1, 2013, Article 2.

[3] M.A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes and M. Slaney, "Content-based music information retrieval: current directions and future challenges," *Proc. IEEE*, 2008, pp. 668–696.

[4] K. Itoyama, M. goto, K. Komatani, T. Ogata and H.G. Okuno, "Query-by-example music information retrieval by score-informed source separation and remixing technologies," *EURASIP journal on Advances in Signal Processing*, Vol. 2010, 2011, Article ID 172961.

[5] Y. Yu, R. Zimmermann and Y. Wang, "Recognition and summarization of chord progressions and their application to music information retrieval," *Proc. IEEE Int. Sympo. Multimedia*, Irvine, California, USA, 2012, pp. 9–16.

[6] L. Lu, d. Liu and H.J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Trans. Audio, Speech and Language Processing*, Vol. 14, No. 1, 2006, pp. 5-18.

[7] D.N. Jiang, L. Lu, H.J. Zhang, J.H. Tao, and L.H. Cai, "Music type classification by spectral contrast features," *Proc. Int. Conf. Multimedia Expo.*, Vol. 1, 2002, pp. 113116.

[8] F. Hitchcock, "The distribution of a product from several sources to numerous localities, " *Journal of Mathematics and Physics*, Vol. 20, 1941, pp. 224–230.

[9] B. Logan and A. Salomon, "A Content-Based Music Similarity Function, " *Tech. Report, Cambridge Research Laboratory*, 2001,

[10] R. Typke, P. Giannopoulos, R. C. Veltkamp, F. Wiering and R. van Oostrum, "Using transportation distances for measuring melodic similarity, " *Proc. Int. Conf. Music Information Retrieval*, 2003, 107–114.

[11] G. Tzanetakis and P. Cook "Musical genre classification of audio signals," *IEEE Trans. Speech and Audio Processing*, Vol. 10, Issue 5, 2002, 293-302.