An Efficient Methodology for the Identification of Multiple Music Works within a Single Query*

Emanuele Di Buccio[1], Nicola Montecchio[1], and Nicola Orio[2]

[1] Department of Information Engineering, University of Padova, Padova IT
[2] Department of Cultural Heritage, University of Padova, Padova IT
{emanuele.dibuccio,nicola.montecchio,nicola.orio}@dei.unipd.it

Abstract. A comprehensive methodology for automatic music identification is presented. The main application of the proposed approach is to provide tools to enrich and validate the descriptors of recordings digitized by a sound archive institution. Experimentation has been carried out on a collection of digitized vinyl discs, although the methodology is not linked to a particular recording carrier. Automatic identification allows a music digital library to retrieve metadata about music works even if the information was incomplete or missing at the time of the acquisition. Results show that the approach is both efficient and effective.

Keywords: Music Identification, Audio Indexing, Cultural Heritage

1 Introduction

In this paper we present a complete methodology for music identification. The main application that motivates our approach is the automatic identification of recordings digitized by sound archives. The experimentation is carried out on a collection of vinyl discs, although the methodology is not linked to a particular carrier and can be readily extended to audio tapes, shellack discs, and so on. Automatic identification allows a music digital library system to retrieve relevant metadata about music works even if this information was incomplete or missing at the time of the digital acquisition. We believe that the availability of systems for the identification of music works will promote the dissemination of music cultural heritage, allowing final users to retrieve and to access individual recordings also when descriptive metadata were not available at the time of acquisition.

A typical approach to automatic music identification is based on the extraction of an *audio fingerprint* from digital recordings. A fingerprint is a compact set of music features that allows for the identification of digital copies even in the presence of noise, distortion, and compression; it can be seen as a content-based signature that summarizes an audio recording [1].

The more general task of music identification differs from audio fingerprint because of the variability of timbre, instrumentation and tempo. Identification of *different versions of the same work* is often based on the extraction of a sequence of audio features

* Extended Abstract

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6,

that are related to high level characteristics of music works – melody, harmony, rhythm – and their alignment using well-known techniques such as Hidden Markov Models (HMMs) and Dynamic Time Warping (DTW) [2].

Identification methodologies based on alignment are usually computationally intensive, because they require the alignment of a query with every item in the reference collection. Quantization of audio features is proposed in [3] to obtain an index-based matching procedure, aimed at efficient music identification. Results showed that the approach is robust to typical variations due to different interpretation of classical music pieces. Our approach, described in detail in [4], is also based on an index data structure to provide an efficient identification methodology.

2 Identification Methodology

Our approach to cover identification aims primarily at efficiency. The methodology is inspired by the concept of Locality Sensitive Hashing (LSH) [5], which is a general approach to handle high dimensional spaces by using ad-hoc hashing functions to create collisions between vectors that are close in the high dimensional space. LSH has been applied to efficient search in media collections [6].

In the proposed approach, the problem of music identification is addressed making use of a particularly compact representation of the audio content, described in Section 2.1. This representation allows the similarity among recordings to be computed by means of a *bag of features* representation, which allows us to exploit indexing techniques to speed up retrieval: Section 2.2 describes how the resulting audio recording representations are stored in an inverted index and the rationale underlying the identification process. Section 2.3 deals with the identification of multiple recordings within a single "long" query.

2.1 Audio Content Representation

A recording is stored in a digital system as a discrete signal, representing an acoustic pressure measured at regular time intervals. In the proposed approach, this *audio waveform* is divided into short overlapping excerpts of fixed length (dozens of milliseconds), and content-based descriptors are then extracted from each resulting segment. The adopted descriptors are *chroma features* [7], 12-dimensional vectors representing the energy associated to each pitch class in a short time frame. A pitch class is a set of pitches corresponding to a given note in the chromatic scale. For instance, the pitch class of the note C corresponds to frequencies: 32.7 Hz, 64.4 Hz, 130.8 Hz, and so on.

Chroma extraction is preceded by *tuning frequency adjustment* in order to be robust to adoption of different reference frequencies, and is followed by a *key finding* procedure, which allows us to deal with different versions of the same music work performed in different keys. The choice of key invariant chroma features as content descriptors is based on the consideration that listeners use mostly harmonic and melodic cues to decide whether two recordings are different performances of the same music work. Differences in tonality, tempo and duration do not substantially affect the similarity

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6,

judgment and thus our representation aims at being robust to changes in these music dimensions.

A general formula to compute chroma vectors from a windowed signal $s(t)$ is the following

$$c_i = \sum_{f=32Hz}^{4000Hz} B_i(f) \cdot S(f) \tag{1}$$

where $S(f)$ is a representation of $s(t)$ in the frequency domain and $B_i(f)$ is a bank of bandpass filters, each centered on the semitones belonging to pitch class i and with a bandwidth of a semitone. Usually chroma features are computed only on a limited range of the audible spectrum.

The particular choice of tuning frequency adjustment, key finding, and chroma extraction algorithms is crucial to the identification accuracy; however, since the focus of the paper is on the identification methodology, rather than signal analysis of music content, we refer the reader to [4] for detailed descriptions of the mentioned algorithms.

Once audio recordings have been represented by chroma vectors, a *hashing* step is introduced to allow for a compact representation of chroma vectors. The quantized version q of a chroma vector $\mathbf{c}$ is obtained by taking into account the ranks of the chroma pitch classes, sorted by their values. Let r_i be the position in which the i-th component c_i would be ranked after a descending sort (starting from 0); a k-level rank-representation of $\mathbf{c}$ is constructed by considering a base 12 number computed as:

$$r_i = |\{c_j : c_j > c_i, j = 1, \ldots, 12\}| \qquad i = 1, \ldots, 12 \tag{2}$$

$$q = \sum_{i:r_i<k} i \cdot 12^{r_i} \tag{3}$$

2.2 Efficient Identification

The identification algorithm is based on a two-level bag-of-features representation of the audio content: the hash sequence computed from the audio waveform according to the procedure described above is divided into overlapping segments of fixed length (typically corresponding to dozens of seconds). Identification of an unknown recording is carried out computing its similarity with the recordings in the collection; adopting the textual IR terminology, we will refer to them as *query* and *documents*, respectively.

Let Q (D) denote a recording associated to a query (audio document in the database), composed of a sequence $\{q_1 \ldots q_{|Q|}\}$ ($\{d_1 \ldots d_{|D|}\}$) of hash sequences, each of length $|q|$ ($|d|$). It is possible to conceptually distinguish two phases in the similarity computation procedure. The first step implies the computation of local similarity between segments d and q as the (normalized) number of descriptors they have in common

$$S_L(d, q) = \sum_{t \in q \cap d} \min\Big(\frac{\mathrm{tf}(t, d)}{|d|}, \frac{\mathrm{tf}(t, q)}{|q|}\Big) \tag{4}$$

where $\mathrm{tf}(t, d)$ denotes the count (*term frequency*) of hashes with value t in segment d. The second step aggregates the contributions of all the query segments, by computing

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6, Copyright (c) 2012 - Edizioni Libreria Progetto and the authors

the geometric mean of the best local similarity values for each query segment:

$$S(Q, D) = \sqrt[|sQ|]{\prod_{q \in Q} \max_{d \in D} S_L(d, q)} \quad (5)$$

An efficient implementation of the similarity computation procedure is possible because any information regarding the *ordering* of segments, and of the hashes contained therein, is discarded. This is reflected in the notation of Equation 5 by the exclusive use of *set operations*.

Data Representation The computation of the similarity score for a query-document pair requires information on the hash frequency in each query and document segment. Information on the frequency of occurrence of an hash in a specific query segment can be extracted at query time and efficiently accessed by means of data structure maintained in memory. The current implementation of the architecture exploits a list of maps, where each list entry (each map) corresponds to a segment and retains *(hash,frequency)* pairs.

Indexing Information on the frequency of a descriptor in a document can be efficiently accessed by means of an inverted index. In such data structure an *inverted list* is associated to each distinct descriptor appearing in at least one of the documents in the collection; the entries of the inverted list are the (identifiers of the) documents where the descriptor occurs. Moreover, additional information necessary by the ranking function can be stored in the inverted list entries, e.g. the frequency of occurrence of the descriptors or the positions where they occur.

In the adopted methodology, the descriptors are chroma hashes, and each segment of a recording is interpreted as a textual document. Figure 1 provides an example of the indexing process when applied to a recording, represented as a sequence of hashes. Such sequence undergoes a segmentation process where possibly overlapping subsequences are extracted from the recording sequence. After segmentation, a recording is represented by a *set* of hash sequences. In accordance with the *bag of features* paradigm, it is hypothesized that information on hash occurrences at a segment level is sufficient for effectively identify a recording; positional information is therefore ignored, thus each segment is effectively treated as a *set* of hashes. An inverted index can efficiently store hash occurrence information: the list of index descriptors is the set of distinct hashes in all the recordings in the collection, while the entries in the inverted list for a hash are the *(segment,frequency)* pairs for that hash.

Document At A Time Processing The adoption of an inverted index allows us to compute efficiently the segment similarity score in Equation 4. When a query Q is submitted to the system it undergoes the same steps as the documents in the collection. The key finding algorithm mentioned in Section 2.1 is applied to the query, thus obtaining diverse transpositions that the system treats as distinct queries, processed in parallel. After the computation of its hash values, the query is split into a number of *segment-queries*. Each segment-query is processed by a Document At A Time (DAAT) strategy. DAAT strategies evaluate the contributions of every query term with respect to a single document before considering the next document. One advantage of the DAAT strategy is

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6,

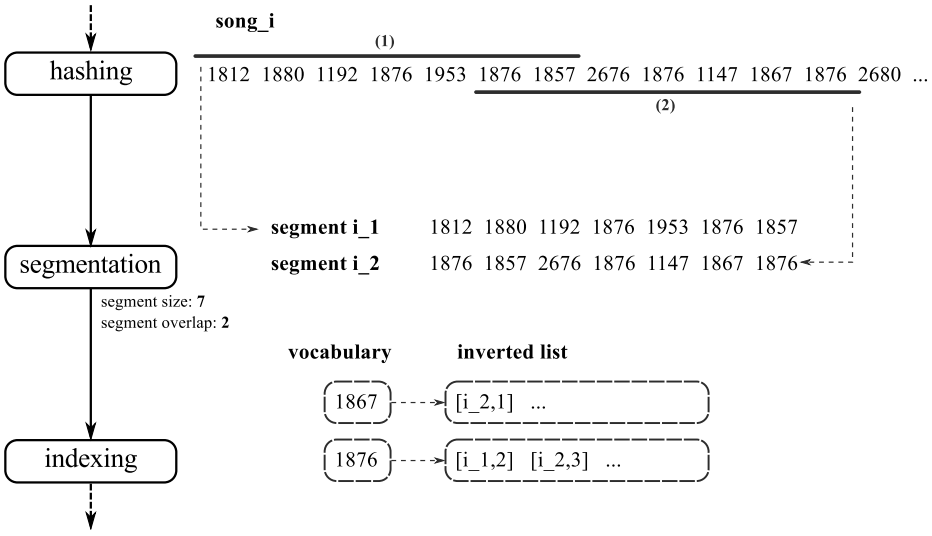

Fig. 1. Segmentation and indexing of hashed chroma descriptors.

that it does not require intermediate document scores to be maintained during the entire ranking process, thus limiting the run-time memory usage.

Score Fusion For each segment-query the system returns a ranked list whose entries are the document segments of all the documents in the collection ranked in decreasing order of similarity value computed at the segment level. The maximum among all the document segments is then computed by going through the returned ranked list. The next step consists in the aggregation of the results returned by the segment-queries, according to Equation 5. Finally, the results obtained from the diverse considered transpositions are merged to obtain a final results list. An advantage of this retrieval strategy is that the diverse query (transpositions) and the diverse segment-queries can be processed in parallel.

2.3 Identification of Multiple Works Within a Single Query

The algorithm detailed above assumes that a query can be matched with a recording of the same music material. This is the typical scenario for music identification systems, but in order to handle the case of a query containing multiple works – such as digitizations of tapes or vinyl discs containing several tracks – the methodology must be extended.

To this end, an additional time-resolution level was added to the segmentation hierarchy. Basically, the recording of an LP side is divided into shorter elements, each one considered as an individual query. Figure 2 shows the procedure, which provides us with a number of resulting rank lists that are merged into a structure called *rank ma-*

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6,

trix. We will refer to these short elements as "chunks", to disambiguate them from the segments in which a single track is divided.

The rationale behind this choice is related to the way the similarity between a query and the documents is computed. As can be seen from Equation 5, it involves a maximization over indexed segments of the local similarities for each query segment. As a consequence, the system is designed to support short queries containing a portion of a corresponding recording while it becomes ineffective with long queries containing segments of different recordings, because the geometric mean will be computed also from local scores with very low similarity values. It should be noticed that, while the maximization and averaging indexes of Equation 5 – d and q respectively – could be in principle reversed (there is no theoretical reason for the asymmetry of the global similarity function), several implementation choices aimed at efficiency depend on this configuration.

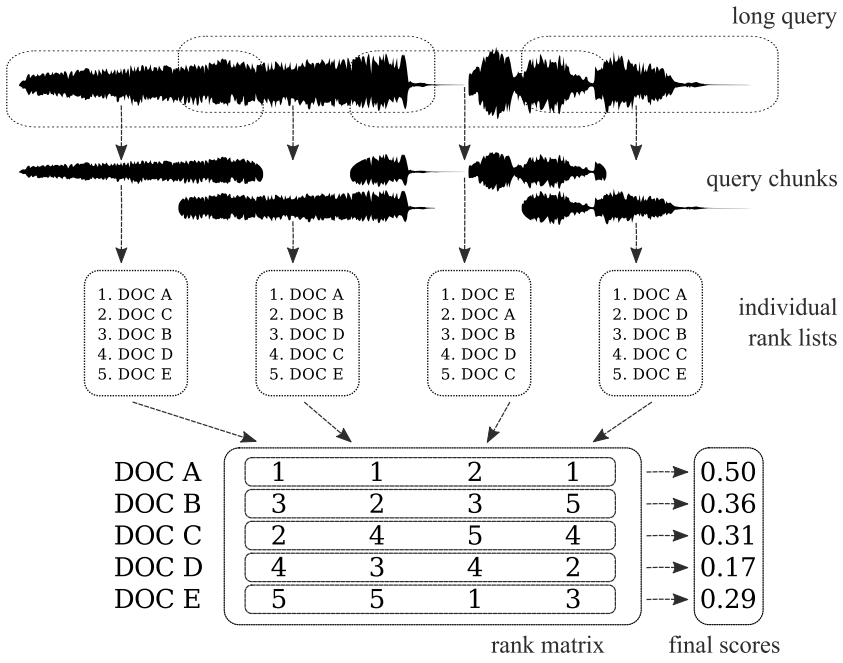

Fig. 2. Computation of the rank matrix for chunks of a long query.

The similarity S_i, between the query and the recording indexed in the i-th row of the rank matrix, is computed with a simple data fusion approach:

$$S_i = \max_{w=1\ldots L-W} \sum_{j=w}^{w+W-1} \frac{1}{(\max(r_{i,j}, C))^2} \tag{6}$$

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6, Copyright (c) 2012 - Edizioni Libreria Progetto and the authors

where L denotes the number of query chunks (columns of the rank matrix), and $r_{i,j}$ represents the position of recording i in the rank list produced from chunk j. The underlying idea is to consider each row of the rank matrix, and to analyze small windows of length W: an indexed recording spanning multiple chunks is supposed to consistently rank in high positions, and the windowing (along with the saturation constant C) acts as a filter for documents whose behavior in the ranking sequence is noisy.

The choice of using rank values instead of similarity scores is motivated by the consideration that averaging similarity scores works only if all segments belong to the same recording, whereas different recordings induce radically different similarity values. As for many data fusion approaches, the choice of rank values reduces the effect of large differences in the similarity scores. Moreover, the choice of ranks enables independence of the particular parametrization adopted for the local similarity computation.

3 Experimental Evaluation

The objective of the experiments reported in this section is to evaluate the capability of the proposed methodology to identify multiple works within a music recording. In particular, it should be verified whether a fixed-length segmentation strategy is able to support identification of recordings characterized by radically different durations.

3.1 Test Collection

The adopted experimental collection is composed of two corpora. The first is part of the test collection that was recently introduced and adopted in the Music Identification Task of the MusiCLEF Lab in CLEF2011 [8]. It is made up of 6680 tracks, grabbed from commercial CDs; among those, 2671 music works are represented at least twice in the collection, forming 945 cover sets. The second corpus is a collection provided by the Fonoteca of the University of Alicante and constituted of 100 LPs that were digitized using common LP record player equipment. The LPs from the Fonoteca were used to evaluate the procedure detailed in Section 2.3, matching their contents with recordings in the MusiCLEF2011 collection.

3.2 Results

The accuracy of the identification methodology is greatly affected by the choice of a particular combination of parameters, among which are the chroma extraction and the tuning frequency adjustment algorithms, and the values of the quantization and segmentation parameters. In order to select an effective combination of parameters, we used the reference recording dataset of the MusiCLEF2011 collection. Considering the accuracy/efficiency trade off, we selected the chroma feature extraction algorithm proposed in [9], using segments of 50s without overlap and 3-level chroma quantization. The results obtained for all the different configurations are reported and discussed in [4].

The chosen parametrization was subsequently used to perform the LP identification task, where multiple works are present within a single query. The effect of different

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6,

Table 1. Experimental results, on a database of 6680 recordings, using a 3.4 Ghz machine.

Task	MRR	MAP	mean query time	mean query length
Track identification	.832	.766	2s	5'25"
LP identification	.900	.742	46s	47'13"

choices for the segmentation and length of the analysis window on accuracy and efficiency of the algorithm was also investigated.

Table 1 reports the accuracy of the algorithm in terms of Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP), along with the average query time (in the LP identification case, the average query time refers to the time needed to identify a complete LP). The experiments were performed on a machine featuring a dual-core 3.4 Ghz CPU and a 7200 RPM hard disk.

The proposed approach is both effective and efficient. In particular, identification can be carried out in a small fraction of the time required for ripping a track from a commercial CD and, most of all, digitizing a complete LP.

References

1. Cano, P., Batlle, E., Kalker, T., Haitsma, J.: A review of audio fingerprinting. Journal of VLSI Signal Processing **41** (2005) 271–284
2. Serrà, J.: Identification of versions of the same musical composition by processing audio descriptions. PhD thesis, Universitat Pompeu Fabra, Barcelona (2011)
3. Kurth, F., Müller, M.: Efficient index-based audio matching. IEEE Transactions on Audio, Speech, and Language Processing **16**(2) (2008) 382–395
4. Montecchio, N., Di Buccio, E., Orio, N.: An efficient identification methodology for improved access to music heritage collections. Journal of Multimedia **7**(2) (2012) 145–158
5. Gionis, A., Indyk, P., Motwani, R.: Similarity search in high dimensions via hashing. In: Proceedings of the International Conference on Very Large Data Bases, Edinburgh, UK (1999)
6. Slaney, M., Casey, M.: Locality-sensitive hashing for finding nearest neighbors [lecture notes]. Signal Processing Magazine, IEEE **25**(2) (2008) 128–131
7. Fujishima, T.: Realtime chord recognition of musical sound: a system using common lisp music. In: Proceedings of the International Computer Music Conference, Beijing, China (1999)
8. Orio, N., Miotto, R., Montecchio, N., Rizo, D., Lartillot, O., Schedl, M.: Musiclef: a benchmark activity in multimodal music information retrieval. In: Proceedings of the Internation Society for Music Information Retrieval conference, Miami, FL, USA (2011)
9. Lartillot, O.: A comprehensive and modular framework for audio content extraction, aimed at research, pedagogy, and digital library management. In: Proceedings of the Audio Engineering Society convention, London, UK (2011)

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6,