Integration of Chroma and Rhythm Histogram Features in a Music Identification System

Riccardo Miotto and Nicola Montecchio
Department of Information Engineering
University of Padova
Padova, Italy
{riccardo.miotto,nicola.montecchio}@dei.unipd.it

***Abstract*—A system for Music Identification is presented which makes use of two different feature descriptors, namely Chroma and Rhythm Histogram features. A simple approach to feature combination is proposed and evaluated against a test collection. Finally directions for future work are proposed, focusing on performance optimization towards a scalable system.**

I. Introduction

The increasing availability of large music collections poses challenging research problems regarding the organization of documents according to some sense of similarity. Following a peculiar social phenomenon in the last years, an increasing number of users is joining social communities to upload their personal recordings and performances. Content-based music identification has become an important research topic because it can provide tools to efficiently retrieve and organize music documents. In particular, the large availability of non-commercial recordings puts a major interest towards the *cover identification* problem. Generally, the term *cover* defines a new rendition of a previously recorded song in genres such as rock and pop. Cover songs can be either live or studio recordings and may have a completely different arrangement.

An earlier approach to music identification was *audio fingerprinting*, that consists in a content-based signature of a music recording to describe digital music even in presence of noise, distortion, and compression [1]. On the contrary, cover identification approaches must be able to identify a song from the recording of a performance, yet independently from the particular performance. For example, identification of live performances may not benefit from the fingerprint of other performances, because most of the acoustic parameters may be different. Collecting all the possible live and cover versions of a music work is clearly unfeasible.

Cover music identification methodologies described in literature generally exploit the well-known *Chroma* features to describe the harmonic content of the music recordings. In particular they have been widely exploited in [2], [3] and [4].

Since Chroma features are high dimensional, they considerably affect computational time for search operations, especially considering the management of different tempo with alignment techniques. Efficiency becomes then a key issue if an identification system is proposed to a large community of users, as in the case of a Web-based music search engine.

In [4], we proposed an efficient methodology to identify classical music recordings by applying the Locality Sensitive Hashing (LSH) paradigm [5], a general approach to handle high dimensional spaces by using ad-hoc hashing functions to create collisions between vectors that are close in the space. LSH has been applied to efficient search of different media [6].

The focus of this paper is on the integration of feature descriptors that are relative to two characterizing aspects of a song: *harmonic* and *rhythmic* content. The main idea is that usually cover songs preserve not only the harmonic-melodic characteristics of the original work but also its rhythmic profile. Starting from the idea proposed in [4], we propose a cover identification system which combines the efficient hash-based Chroma descriptors with a rhythmic profile descriptor in order to increase the identification precision.

II. System model

The system works by combining the evidence given by Chroma and Rhythmic descriptors into a single ranking of possible matching songs. The rhythm descriptors used are *Rhythm Histogram* (RH) features [7], which were originally proposed in a genre classification task.

While Chroma features have been thoroughly investigated previously and are provided with an efficient implementation, RH features have only recently been adopted by us and their performances in terms of speed are not comparable yet; both aspects are described below. An overview of the system is depicted in Figure 1.

A. Chroma features

A music descriptor widely applied to cover identification is Chroma features. As is well-known, Chroma features are related to the intensity associated with each of the 12 semitones within an octave, with all octaves folded together. The concept behind chroma is that the perceived quality of a chord depends only partially on the octaves in which the individual notes are played. Instead, what seems to be relevant are the pitch classes of the notes (the names of the notes on the chromatic scale) that form a chord. This robustness to changes in octaves is also exploited by artists who play a cover song: while the main melody is usually very similar to the original one, the accompaniment can have large variations without affecting the recognizability of the song.

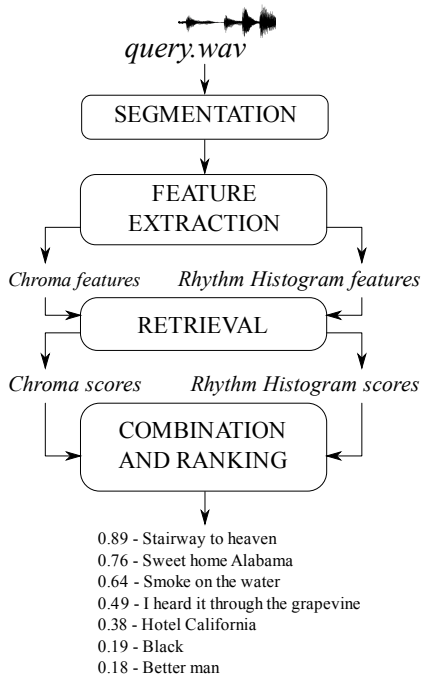


Fig. 1. Overview of the system model

As described in [4], a Chroma vector $\mathbf{c}$ is a 12-dimensional vector of pitch classes, computed by processing a windowed signal with a Fourier transform. According to the approach proposed in [8], chroma features have been computed using the instantaneous frequency within each FFT bin to identify strong tonal components and to achieve higher resolution. In Figure 2(a) a Chroma vector corresponding to an A7 chord is depicted; Figure 2(b) shows the evolution of Chroma vectors over time for an excerpt of the song "Heroes" by D. Bowie.

For each vector $\mathbf{c}$, quantization $\mathbf{q}$ is achieved by considering the ranks of the chroma pitch classes, instead of their absolute values, to obtain a general representation robust to variations due to different performing styles. In particular, rank-based quantization is carried out by computing the rank of the value of the energy in the various pitch classes.

Rank-based quantization aims at a final compact representation, which can be obtained by considering that a vector $\mathbf{q}$ can be thought as a twelve digit number represented in base k. A simple hashing function h can be computed by obtaining the decimal representation of this number, according to equation

$$h = \sum_{i=1}^{12} k^{i-1} \mathbf{q}_i \qquad (1)$$

where additional hashing techniques can be applied to store the values h in one array, which can be accessed in constant time. A typical technique is to compute the reminder of h divided by a carefully chosen prime number.

The described approach is applied both to the songs in the collection and to the queries. With the main goal of efficiency, retrieval is carried out using the *bag of words* paradigm. Similarity between the query and the recordings in the collection is measured by simply counting the number of hashes they have in common. This measure of similarity does not take into account the distribution of hash values. In particular, the occurrence of a chroma hash inside a song and the frequency of a chroma hash across the collection of documents have not been considered. The choice of this particular similarity measure has been motivated by a number of tests using short queries of about 10 seconds, where this simple measure outperformed more complex ones [4].

Since queries of a cover identification task can be complete recordings, the frequency of chroma hashes and their relative position along the song may become a relevant piece of information. In order to handle this issue, long music queries have been divided in overlapping short sub-queries of a fixed duration and for each query an independent retrieval task is carried out. A similar processing is applied to documents, that are divided in overlapping frames with a length comparable to the one of the sub-queries. At the end, the results score of each single retrieval are combined. In particular, preliminary evaluation showed that, in this context, geometric mean outperformed all the other main fusion techniques reported in literature [9].

A problem that may affect retrieval effectiveness is that chroma-based representation is sensible to transpositions. The problem is not dealt with in this paper, as the focus mainly resides in the integration with rhythmic features; it is however part of future work and possible solutions are described in Section IV.

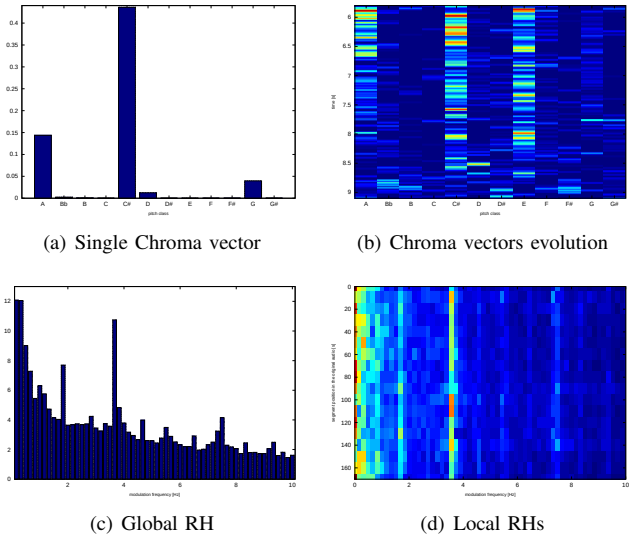
B. Rhythm Histogram features

Rhythm Histogram features [7] are a descriptor for the general rhythmic characteristics of an audio document. In a RH the magnitudes of each modulation frequency bin for all the critical bands of the human auditory range are summed up to form a histogram of "rhythmic energy" per modulation frequency.

In their original form, a single "global" RH represents a whole piece; in our approach, as is the case for Chroma features, a sequence of RHs is computed for each song by segmenting the audio into overlapping windows of 15 seconds, in order to be able to individually match parts of songs which might be characterized by different rhythmic structures (e.g. verse and chorus). Figures 2(c) and 2(d) show the global RH and the sequence of local RHs for David Bowie's "Heroes".

The first step in the computation of the similarity between the songs a and b is the construction of the similarity matrix M, in which each entry m_{ij} is given by the cosine similarity of the i-th RH of a and the j-th RH of b. For each segment of a, the best matching segment of b (that is the one with the highest cosine similarity) is retained, and the mean of these values over all segments of a is computed; a simmetric procedure is then applied to song b and finally the average[1] of these two

[1]A bi-directional comparison procedure is used in order to have a symmetric similarity measure. Experiments however showed that the simpler unidirectional comparison strategy yields similar results.



(a) Single Chroma vector (b) Chroma vectors evolution

(c) Global RH (d) Local RHs

Fig. 2. Chroma and Rhythm Histogram features for D. Bowie's "Heroes"

scores is returned as the similarity of a and b. Experimental results showed that this strategy performs slightly better than the simpler comparison of the global RHs.

It is clear that this approach is computationally intensive, since the cosine similarity of the RHs must be computed for each song in the collection and for each segment pair. Possible optimizations, similar to the ones used for Chroma features, are under investigation. In Section III-C a straightforward strategy for reducing the computational load is proposed, based on the consideration that query songs can be compared to just a small subset of the songs in the collection while retaining the same precision in the results.

C. Feature combination

The similarity score s for a pair of songs, that governs the ranking returned by the system, is computed combining the two scores c and r given by the Chroma features and the Rhythm Histogram features respectively. Two strategies have been used:

- linear combination

$$s = (1-\alpha)c + \alpha r \qquad \alpha \in [0,1] \tag{2}$$

- weighted product

$$s = c^{1-\alpha} r^{\alpha} \qquad \alpha \in [0,1] \tag{3}$$

As pointed out in Section III-C, their performance is similar.

III. Experimental results

Experimental results are presented to show how performances can be improved by combining the scores for the two feature descriptors used. The performances of the system are evaluated using Mean Reciprocal Rank (MRR) as a measure of precision.

A. Test collection

The proposed approach has been evaluated with a test collection of 500 recordings of pop and rock songs, taken from personal collections of the authors. The idea was to have a collection as close as possible to a real scenario. In fact, the indexed documents were all the original version of the music works – i.e., the studio album versions – for which it is expected that metadata are correctly stored and that can be reasonably used in a real system as the reference collection.

The query set included 60 recordings of different versions of a subset of the collection, which were live version by the same artist who recorded the original song and studio or live covers by other artists. We decided to have in the collection one single correct match for each query in order to balance the contribution of each different song. Queries had different durations, generally including the core of the music works – verses and choruses – plus possible introductions and endings, in particular in the live versions. All the audio files were polyphonic recordings with a sampling rate of 44.1 kHz and stored in MP3 format at different bitrates (at most 192 kbps). In fact, in order to simulate a real context, we preferred a compressed format rather than an uncompressed one such as PCM.

B. Individual features results

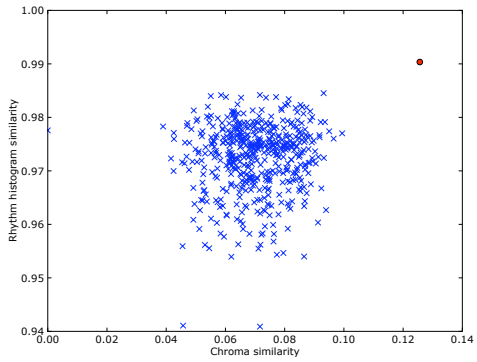
The performance of Chroma features individually is already satisfying, with a MRR of 78.4%. Rhythmic Histogram features on the other hand are less reliable, resulting in a MRR of 34.0%. If the RH features scores are computed directly on the global RH (instead of subdividing the song and computing the best match for each segment) MRR is is 28.5%.

C. Combined features results

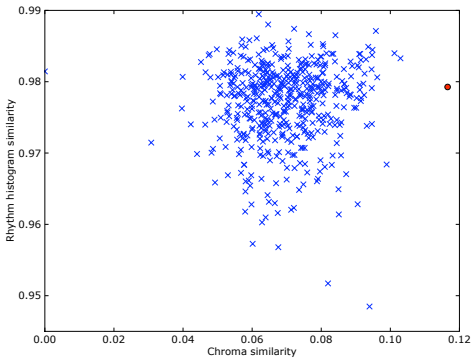
Figure 3 shows typical dispositions of the feature score pairs for some queries; each point in the feature space is associated to a comparison between a query song and the songs in the collection, the red circles being associated to the relevant matches. In particular Figure 3(a) is an example of the best possible situation, in which both Chroma features and Rhythm Histogram features individually rank the correct match for the query song in the first position. Figure 3(b) depicts the most common situation, in which Chroma features correctly identify the match but RH features are misleading; the dual situation is reported in Figure 3(c), which is rare but represents a significant evidence for the usefulness of RH features. Finally Figure 3(d) presents the situation in which neither feature descriptor can correctly identify the best match.

Perhaps the most interesting disposition of score pairs in the feature space is the one depicted in Figure 4: neither feature can identify the matching song by itself, but a combination of the two is indeed able to rank it in the first position.

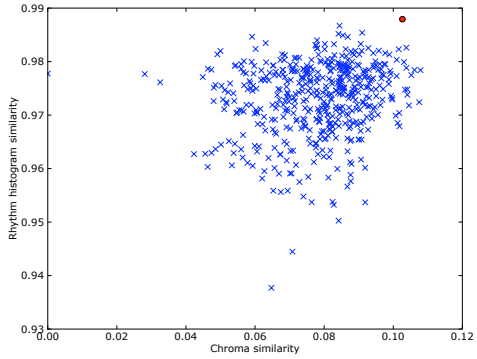
The two approaches to feature combination reported in Equations 2 and 3 have been tested for several values of the parameter α, which weights the influence of RH features in the score, and the resulting MRRs are depicted in Figure 5. For the optimal value of α, the MRR increases from 78.4%



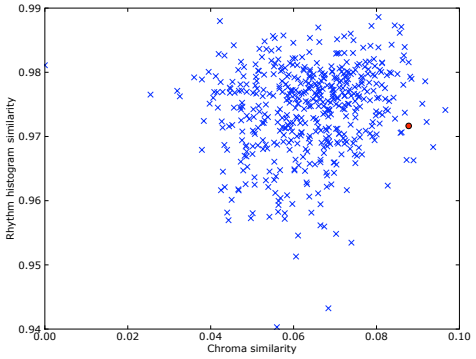
(a) Smoke on the water - Deep Purple



(b) Sweet child of mine - Guns N' Roses



(c) Sweet home Alabama - Lynyrd Skynyrd



(d) You shook me - AC/DC

Fig. 3. Disposition of song similarities in the feature space

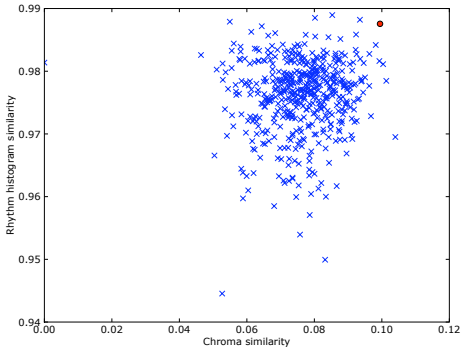


Fig. 4. Disposition of song similarities in the feature space for "All along the watchtower" by Jimi Hendrix

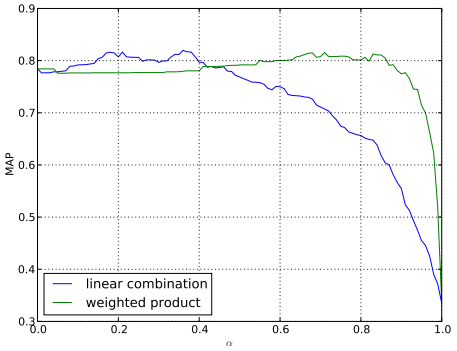


Fig. 5. MRR for the presented approaches to feature combination

(using only Chroma features) to 82.0% and 81.6%, using a linear combination and a weighted product of the features scores respectively. Similar performances are achieved using a single global RH for computing the similarity of songs, with a MRR of 81.5% for the case of the best linear combination of features. Even though the MRR maximum value is located in local peaks of the graphic, which are probably due to the rather small size of the test collection, setting α in a rather large neighbourhood of its optimal value still yields a significant improvement in MRR.

As anticipated in Section II-B, it is clear that performing the comparison of a query song against the whole set of songs in the collection is unfeasible for a large collection,

especially when comparing all the segments against each other. Fortunately Chroma features are able to rank the relevant match in the first positions, and this can be done efficiently thanks to the hashing mechanisms discussed above; an effective solution is to exploit this robustness by reranking only the top t position with the aid of Rhythm Histogram features: with t ranging from 15 to 50 the optimal MRR (82.0%) is unchanged for the collection used. Although the collection is very small, previous experiments with Chroma features on larger collections [4] have shown how the relevant matches for query songs are almost never ranked in very low positions, thus Rhythm Histogram features can be effectively exploited computing them on just a very small fraction of the songs in the collection.

IV. Conclusion

A system for cover identification of pop songs has been presented, focusing on the improvement in identification precision given by the introduction of a rhythmic profile descriptor in addition to the modeling of harmonic content. Many directions for future work are yet to be explored, and the most promising ones are briefly reviewed below.

The modeling of harmonic content still lacks an effective solution for handling the possible transpositions in pitch of cover songs; in fact, if the cover song used as query and the original work stored in the collection are played in different tonalities, they have totally different sets of chroma. This problem can be addressed by considering that a transposition of n semitones will result in a rotation of Chroma vectors of n steps. Then, the tonality issue can be dealt with by simply computing all the twelve transpositions, but at the cost of a loss in precision. Alternatively, a methodology for key finding can be exploited to compute the similarity between the songs in the collection by transposing their chroma features into a reference tonality [3]. The combination of key finding algorithms with the proposed approach will be an important part of future work.

Our investigation of rhythmic content descriptors is still in an early stage. Computational performances are a primary concern, and an efficient implementation of RH retrieval similar to the hash-based implementation of Chroma retrieval is under exhamination.

Finally it is clear how evaluation of the system should be performed on a bigger test collection; this however poses additional issues, not only related to the size of the data that has to be managed, but also to problems regarding music genres, which might have to be dealt specifically: in particular many genres are defined by a very characteristic rhythm (e.g. reggae) thus rhythmic descriptors might be in such cases detrimental to the final performances.

References

[1] P. Cano, M. Koppenberger, and N. Wack, "Content-based music audio recommendation," in *Proceedings of the ACM International Conference on Multimedia*, 2005, pp. 211–212.

[2] F. Kurth and M. Muler, "Efficient index-based audio matching," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 2, pp. 382–395, 2008.

[3] J. Serra, E. Gomez, P. Herrera, and X. Serra, "Chroma binary similarity and local alignment applied to cover song identification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 6, pp. 1138–1151, 2008.

[4] R. Miotto and N. Orio, "A music identification system based on chroma indexing and statistical modeling," in *Proceedings of International Conference on Music Information Retrieval*, 2008, pp. 301–306.

[5] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in *The VLDB Journal*, 1999, pp. 518–529.

[6] M. Slaney and M. Casey, "Locality-sensitive hashing for finding nearest neighbors [lecture notes]," *Signal Processing Magazine, IEEE*, vol. 25, no. 2, pp. 128–131, 2008.

[7] T. Lidy and A. Rauber, "Evaluation of feature extractors and psychoacoustic transformations for music genre classification," in *Proceedings of International Conference on Music Information Retrieval*, 2005, pp. 34–41.

[8] D. Ellis and G. Poliner, "Identifying 'cover songs' with chroma features and dynamic programming beat tracking," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, April 2007, pp. IV–1429–IV–1432.

[9] E. Fox and J. Shaw, "Combination of multiple searches," in *Proceedings of the Second Text REtrieval Conference (TREC-2)*, 1994, pp. 243–249.