A HEURISTIC FOR DISTANCE FUSION IN COVER SONG IDENTIFICATION

Alessio Degani, Marco Dalai, Riccardo Leonardi and Pierangelo Migliorati

University of Brescia
DII, Signals and Communication Lab
Via Branze, 38, 25123 Brescia, ITALY

ABSTRACT

In this paper, we propose a method to integrate the results of different cover song identification algorithms into one single measure which, on the average, gives better results than initial algorithms. The fusion of the different distance measures is made by projecting all the measures in a multi-dimensional space, where the dimensionality of this space is the number of the considered distances. In our experiments, we test two distance measures, namely the Dynamic Time Warping and the Q_{max} measure when applied in different combinations to two features, namely a Salience feature and a Harmonic Pitch Class Profile (HPCP). While the HPCP is meant to extract purely harmonic descriptions, in fact, the Salience allows to better discern melodic differences. It is shown that the combination of two or more distance measure improves the overall performance.

Keywords: cover song identification, distance fusion

1. INTRODUCTION

Cover song identification aims at finding different *versions* of the same musical piece within a large database of songs. In the last 10 years, a lot of work has been made to try to successfully accomplish this task. Thanks to the MIREX evaluation campaign for Music Information Retrieval (MIR) algorithms, this research topic has gained attention and methods have improved in accuracy. Different algorithms have been developed in the literature and the standard approach to measuring similarity between cover songs is to exploit music facets shared between them. This similarity measure is computed using different descriptors (or features) extracted from the raw audio file. In order for these descriptors to be effective, they have to be relatively insensitive to the majority of musical changes among covers, like tempo or key. Once the descriptors are extracted, a measure of distance between song descriptions is evaluated and a similarity score between songs is thus obtained. Hence, a cover song identification algorithm usually takes a query song as an input and, after a processing step for the extraction of the descriptors, performs a comparison between this song and all songs in a database, using the extracted feature. The result of this run is a ordered list of songs ranked with a distance criteria, where the most similar song must ideally rank first in this list.

Of course, different features and different distances over such features can be used. The first largely used descriptor, the so called Pitch Class Profile (PCP), was introduced in 1999 by Fujishima [1]. Over the years, PCP (sometimes also called chromagram) was extended and modified in different variants, some of which are still successfully used in cover song identification (see HPCP [2]).

Among the possible techniques to compute distances between sequences of features, we can list two which are of particular importance for the cover song identification problem. One is the Dynamic Time Warping (DTW [3, Chapter 4]), that aims to find an optimal alignment path between two given time-dependent sequences. It gives us a warping cost value between a song u and a song v. Another technique is introduced in [4], which uses the Cross-Recurrence Plot [5] and recurrence quantification analyses like Q_{max} [6].

Different performances are obtained when using a specific feature with a specific distance measure, and it is not always easy to understand which feature-distance combination behaves better, since this obviously depends on the dataset at hand, on the query etc. Moreover, even a single feature, when extracted with different settings, can give different performance. Based on this fact, for example, the system Hydra [7] combines features and distances extracted with different parameters which are fed to a Support Vector Machine which output, for each pair of songs, a single bit decision of the type cover/non-cover. A similar approach is used in [8], where a distance is calculated over three different audio descriptor and a classifier is trained with a subset of known cover or non-cover songs pairs. In our work instead, we do not apply any classification and no training is needed.

In this paper we discuss an approach to cover song identification that involves a blind combination of different features and different distance measures without making any assumption on the audio descriptor used. For the evaluation, we consider two type of features, the Salience function and the HPCP combined with two distance measure, namely Dynamic Time Warp and a Q_{max}. Each feature, when used with a given distance, allows sorting the songs in the database in order of decreasing similarity with a given query song. In this paper,

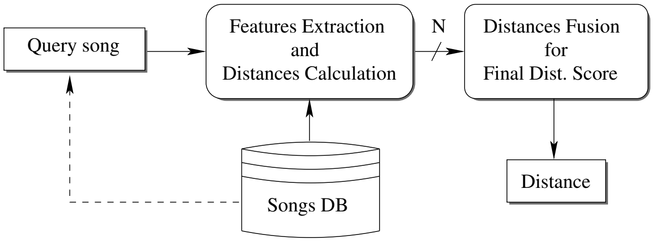
Query song
Features Extraction and Distances Calculation
N
Distances Fusion for Final Dist. Score
Songs DB
Distance

Fig. 1. General scheme of the method

we propose a technique to combine the lists obtained by N feature-distance combinations into one single N-dimensional space in order to assess a "globally informed" distance measure. As we will see in Section 4, the so obtained combined list leads to a relative improvement in accuracy with respect to the results obtained by the single measures separately. In the Figure 1 we can see the general outline of the algorithm.

In the Section 2 we explain what features and what distances are involved in our test and in the Section 3 we show how we combine them. In Section 4 and 5 we present and discuss the accuracy results.

2. AUDIO FEATURES AND DISTANCE METRICS

2.1. Audio Features

Here we give a brief introduction of the audio descriptors used in our test. The described features are computed for each frame for a total of N_f frames. This leads to a huge amount of data which would make the complexity of any distance measure evaluation impractical. Hence, a temporal downsampling is applied to the sequence of features to obtain a shorter sequence of length $N_t < N_f$. In our case, we use an adaptive decimation factor that is dependent on the beat duration in order to obtain a beat-synchronous time average. This part is based on the algorithm presented in [9]. Where not explicitly mentioned all of the feature and distance calculation algorithms are re-implemented by the authors of this paper.

2.1.1. Pitch Salience Function

As presented in [10], a salience function for a given frequency f_i is calculated as a weighted sum of the energy at the first 8 harmonics of the fundamental frequency f_i like f_i, $2f_i$, $3f_i$ $, \dots$. The pitch salience function is calculated at each frame using the amplitude spectrum and covers a frequency ranging from 55 Hz to 1.76 kHz (5 octaves range from $A1$ to $A6$) using a resolution of resolution of 1 bin/semitone.

2.1.2. HPCP

Briefly, the computation of the HPCP descriptor begins with some pre-processing step like spectral peak detection [11]. Next, the energy of each spectral peak from about 50 Hz to 5 kHz is represented as a vector of Pitch Classes energy. Each Pitch Class represents a semitone in a twelve-tone equal-tempered chromatic scale. The energy of each Pitch Class are calculated from the correspondent spectral peak and the weighted summation of its harmonic frequencies peak energy up to 8 terms. The reader is referred to [2] for a complete description.

2.2. Distance Measures

In this section we give a brief description of the distance measures used for the evaluation of our method.

2.2.1. CRP/Qmax

Basically, the Q_{max} distance calculates the length of the longest time segment in which two song u and v exhibit similar feature patterns. This is done by using a cross-recurrence plot. A cross-recurrence plot (CRP) is a binary similarity matrix $\mathcal{C}$ whose elements $c_{i,j}$ are set to 1 when there is a recurrence between the i-th feature vector of song u and the j-th feature vector of song v, and zero otherwise. Here, a recurrence means that the euclidean distance between this two vectors is below a specified threshold. For more details such as the threshold value and the CRP algorithm see [5, 4]. When consecutive feature vectors are similar for a certain amount of frames, a diagonal patterns of ones become visible in CRP. What the Q_{max} algorithm does is to quantify the presence and the length of this diagonal patterns in the CRP using an efficient recurrence quantification analysis [4]. In a nutshell, a cumulative matrix $\mathcal{Q}$ is computed over the elements of $\mathcal{C}$ starting form the element $c_{1,1}$ and counting the elements with value equal to 1 that are aligned in a diagonal way. Finally, the Q_{max} value is calculated as the maximum amplitude of the elements $Q_{i,j}$ of the matrix $\mathcal{Q}$ as

$$Q_{max} = \max\left(Q_{i,j}\right). \tag{1}$$

The Q_{max} measure gives a similarity quantification of two input songs. In our case we need a distance measure that can be calculated as

$$d_{u,v} = \frac{\sqrt{N_t^v}}{Q_{max}}, \tag{2}$$

where N_t^v is the length of the salience function of song v and plays the role of a normalization factor [4].

2.2.2. Dynamic Time Warping

Dynamic Time Warping [3, Chapter 4] is a technique to find an optimal path to align two time sequences. Ideally, the two sequences are warped in a non-linear way to reach the maximum matching between each other. DTW gives itself a measure of distance between two sequences, and it thus be used to assess similarity between two songs [12, 13]. With DTW, we

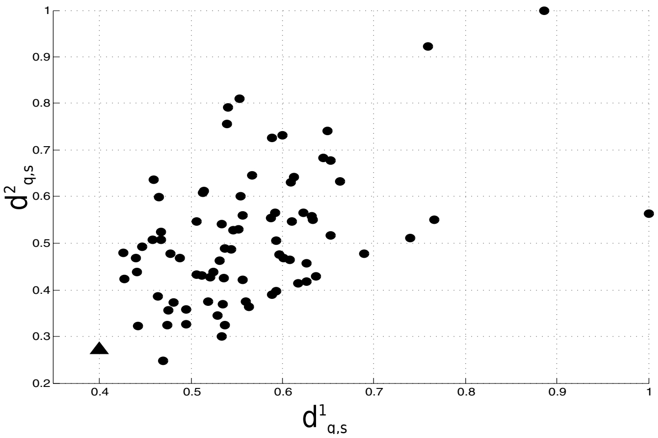

Fig. 2. Cloud for a given $q = q_0$, d^1 calculated using DTW on Salience feature and d^2 with Q_{max} over HPCP. The triangle identifies the correct cover song. Note that in this case the correct cover does not minimize d^2.

obtain the total alignment cost $DTW_{u,v}$ between two features sequences u and v. For more details, see [3, Chapter 4]. We used the DTW implementation freely available at [14], with only one minor modification, namely that the a normalization similar to that of (2) is applied to obtain $d_{u,v}$ as follows

$$d_{u,v} = \frac{DTW_{u,v}}{\sqrt{N_t^v}}. \tag{3}$$

In our tests this normalization leads to a performance improvement.

3. DISTANCE SELECTION

In this Section we describe the proposed technique for the merging of two or more feature and distance measure combinations in order to create a single ranking with improved performance. Independently from the used features and metrics, we assume that N different distance metrics $\mathbf{d}_{q,s} = [d^1_{q,s}, \cdots, d^N_{q,s}]$ are computed between the query song q and each song $s \in [1 \ldots S]$ in the database. In a nutshell, the proposed method mixes N distances by projecting them in a N-dimensional space in order to refine the ranked list in a more reliable way. The process is now described in detail. Assuming the cover song identification algorithm returns a square $S \times S$ cross-distances matrix $\mathcal{D}$ where each element $d_{q,s}$ of this matrix represents a distance between the song q and the song s in the database, and we can calculate more than one distance matrix using different combination of features and metrics, we obtain N distances matrix $\mathcal{D}^1, \cdots, \mathcal{D}^N$. For a fixed query song $q = q_0$ and a fixed distance metric $n = n_0$, we make a normalization of the distance vector as

$$\bar{d}^{n_0}_{q_0,s} = \frac{d^{n_0}_{q_0,s}}{\max\limits_{s \in [1 \ldots S]} [d^{n_0}_{q_0,s}]}, \forall s \in [1 \ldots S] \tag{4}$$

in order to ensure that $\bar{d}^{n_0}_{q_0,s} \in [0, 1]$. Now, for a fixed query $q = q_0$ and a fixed song $s = s_0$ in the database, we define a point in a N-dimensional space that uniquely identifies the position of the pair (q_0, s_0) in the distances space

$$\bar{\mathbf{d}}_{q_0,s_0} = [d^1_{q_0,s_0}, \cdots, d^N_{q_0,s_0}] \in [0, 1]^N. \tag{5}$$

At this point we are able to compute $\bar{\mathbf{d}}_{q,s} \in [0, 1]^N$ for each $q, s \in [1 \ldots S]$. The points $\bar{\mathbf{d}}_{q,s}$ form an N-dimensional cloud. An example of one such cloud with $N = 2$ for a given query q is shown in Fig. 2. We now compute a new $S \times S$ refined distance matrix $\mathcal{R}$ whose elements $r_{q,s}$ are defined as

$$r_{q,s} = ||\mathbf{1}|| - ||\bar{\mathbf{d}}_{q,s} - \mathbf{1}||, \tag{6}$$

where $|| \cdot ||$ is the l_2 norm and $\mathbf{1} = [1, \cdots, 1]$ is the N-dimensional "one" vector. Since the vector $\mathbf{1}$ represent the point in our space where all the distances are maximum, the terms $||\bar{\mathbf{d}}_{q,s} - \mathbf{1}||$ in (6) expresses how far the pair (q, s) is from to be a *non-cover pair*. It follows that (6) can be seen as a measure of how likely the pair is a *cover pair*. The origin of the N-dimensional space is the ideal place where a *cover pair* (q, s) should be placed, so intuitively, one may think that each element of $\mathcal{R}$ can be calculated as $r_{q,s} = ||\bar{\mathbf{d}}_{q,s}||$. Though this approach does work in practice, however, in our tests this strategy leads to worse results compared to (6).

4. RESULTS

The evaluation task is performed using the well known cover song dataset named covers80 [15]. We perform a comparison of the performances between the basic algorithms that use one distance metric over a single feature type, and a number of combinations of choices of features and distance metrics. The used performance indicators are some of the commonly used indicators in Music Information Retrieval: Precision, Mean Rank of First Correctly Identified Cover (MR1st) and Mean Average Precision (MAP). The total virtual score T is calculated by counting the total unique correct identified cover for all of the method in the combination and normalizing by S. As we stated in Section 2, for our evaluation we use two features type: the HPCP (H) and the Salience function (S). For the distance metrics we use the Q_{max} measure (Q) and the dynamic Time Warping (D). Table 1 reports the accuracy indicator for different combination of feature/distance. For the HPCP with Q_{max} approach, the accuracy indicators may be different from the original implementation of the algorithm by J. Serrà [4] since it has been completely rewritten by the authors. As we can see in Table 1, all the accuracy results with $N > 1$ bring an improvement with respect to the simple distance measure. Although the best results are obtained using all ($N = 4$) of the possible combination of basic distances, we can see that using (S,D)+(H,Q) we obtain a comparable result but with a lower dimensionality $N = 2$. Furthermore, we can see that the indicator T is proportional to the MAP. Higher T means higher MAP and consequently better precision.

Combination	N	Prec.	MR1st	MAP	T
(S,D)	1	0.47	12	0.55	-
(S,Q)	1	0.52	15	0.60	-
(H,D)	1	0.35	17	0.43	-
(H,Q)	**1**	**0.59**	**9**	**0.65**	-
(S,D)+(S,Q)	2	0.61	12	0.66	0.61
(S,D)+(H,D)	2	0.46	12	0.53	0.55
(S,D)+(H,Q)	**2**	**0.65**	**7**	**0.71**	**0.66**
(S,Q)+(H,D)	2	0.56	13	0.63	0.60
(S,Q)+(H,Q)	2	0.64	10	0.69	0.69
(H,D)+(H,Q)	2	0.60	9	0.66	0.60
(S,D)+(S,Q)+(H,D)	3	0.60	11	0.66	0.66
(S,D)+(S,Q)+(H,Q)	**3**	**0.65**	**9**	**0.71**	**0.74**
(S,Q)+(H,D)+(H,Q)	**3**	**0.66**	**9**	**0.70**	**0.70**
(S,D)+(H,D)+(H,Q)	3	0.60	7	0.67	0.68
(ALL)	**4**	**0.66**	**8**	**0.72**	**0.75**

Table 1. Accuracy results.

5. CONCLUSIONS

In this paper we presented a method to merge N combination of feature and distance measures to increase the accuracy results of a cover song identification algorithm. This method is based uniquely on a geometric N-dimensional distance measure that has a very low computational cost of the overall distance refinement. A particularly useful combination has been obtained by using a Salience Feature with a Dynamic Time Warp similarity measure and a HPCP with a Q_{max} measure. This combination, in our tests, proved to give excellent performance with a low dimensionality N. The percentage T of the virtual total unique correct identified cover plays a fundamental role for the accuracy performances of the distance fusion process. The most important property of the method, however, is that it can be used to combine any set of different distance metrics, regardless of what they measure and without making any assumption on the specific audio features involved.

6. REFERENCES

[1] Takuya Fujishima, "Realtime chord recognition of musical sound: a system using common lisp music," *Proc. of the Int. Computer Music Conference (ICMC)*, pp. 464–467, 1999.

[2] Emilia Gómez, *Tonal Description of Music Audio Signals*, Ph.D. thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2006.

[3] Meinard Müller, *Information Retrieval for Music and Motion*, Springer, 2007.

[4] Joan Serrà, *Identification of Versions of the Same Musical Composition by Processing Audio Descriptions*, Ph.D. thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2011.

[5] Norbert Marwan, M. Carmen Romano, Marco Thiel, and Jürgen Kurths, "Recurrence plots for the analysis of complex systems," *Physics Reports*, vol. 438, no. 5, pp. 237–329, 2007.

[6] Joan Serrà, Xavier Serra, and Ralph G. Andrzejak, "Cross recurrence quantification for cover song identification," *New Journal of Physics*, vol. 11, 2011.

[7] Suman Ravuri and Daniel P.W. Ellis, "The hydra system of unstructured cover song detection," *MIREX 2009 extended abstract*, 2009.

[8] Justin Salamon, Joan Serrà, and Emilia Gómez, "Melody, bass line, and harmony representations for music version identification," *Proceedings of the 21st international conference companion on World Wide Web (WWW 2012)*, 2012.

[9] Daniel P. W. Ellis, "Beat tracking by dynamic programming," *Journal of New Music Research*, vol. 36, no. 1, pp. 51–60, 2007.

[10] Justin Salamon, Emilia Gómez, and Jordi Bonada, "Sinusoid extraction and salience function design for predominant melody estimation," *Proc. of the 14th Int. Conference on Digital Audio Effects (DAFx-11)*, 2011.

[11] Xavier Amatriain, Jordi Bonada, Alex Loscos, and Xavier Serra, *in Udo Zölzer DAFX-Digital Audio Effects*, chapter Spectral processing, pp. 373–438, John Wiley & Sons, 2002.

[12] Joan Serrà, Emilia Gómez, Perfecto Herrera, and Xavier Serra, "Chroma binary similarity and local alignment applied to cover song identification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16, no. 6, pp. 1138–1152, 2008.

[13] Wei-Ho Tsai, Hung-Ming Yu, and Hsin-Min Wang, "Using the similarity of main melodies to identify cover versions of popular songs for music document retrieval," *Journal of Information Science and Engineering*, vol. 24, pp. 1669–1687, 2008.

[14] Daniel P. W. Ellis, "Dynamic time warp (dtw) in matlab," *URL: http://labrosa.ee.columbia.edu/matlab/dtw/*, 2008.

[15] Daniel P. W. Ellis, "The 'covers80' cover song data set," *URL: http://labrosa.ee.columbia.edu/projects/coversongs/covers80/*, 2007.