THE EXTENDED BALLROOM DATASET

Ugo Marchand & Geoffroy Peeters
STMS IRCAM-CNRS-UPMC
1 pl. Igor Stravinsky, 75004 Paris, France

ABSTRACT

We present here the Extended Ballroom dataset. This dataset is an improved version of the well-known Ballroom dataset. It provides amongst other things, more tracks than the Ballroom dataset and a list of track repetitions (exact duplicates, karaoke versions, ...). Thus it extends the range of possible applications. We describe here how we assembled the Extended Ballroom dataset, how the various annotations (tempo, rhythm class, duplicates, ...) were made and how they can be trusted, as well as the possible applications we can make out of the Extended Ballroom dataset.

1. INTRODUCTION

The Ballroom dataset was created for the rhythm description contest of ISMIR 2004 [1]. It was extracted from the website *www.ballroomdancers.com* [1] at that time. The ballroon test-set contains 698 music exerpts of 30 second each, divided into 8 genres representing various Ballroom dances (ChaChaCha, Jive, Quickstep, ...). As these genres are closely related to rhythmic patterns, they can be considered as rhythm classes.

Due to the relatively low number of tracks, the bad audio quality and the fact that the original website still exists and still offers to listen to 30-second excerpts (along with tempo and genre annotation), we decided to create the Extended Ballroom dataset. We extracted all the audio excerpts from the website, along with all the metadata available. We also annotated semi-automatically all kind of repetitions that can be found among the 4.180 downloaded tracks.

There are multiple advantages to this Extended Ballroom dataset: better audio quality, 6 times more tracks, 5 new rhythm classes and annotations of different types of repetitions (exact duplicate, karaoke version, ...).

[1] This website sells audio CDs of Ballroom dances and offers to listen to a 30-seconds preview of each track.

2. EXTENDED BALLROOM DATASET

In the following we describe this new test-set using the recommendation made by [3] for "the description of annotated MIR corpora". The letters and numbers in brackets (such as '(B31)') refer to the description of [3].

2.1 Audio (A)

We show in Table 1 the new class distribution of the Extended Ballroom dataset which contains now 4.180 tracks. The number of tracks by rhythm class (or genre) went from an average of 87 to 444 tracks (excepting the class ViennneseWaltz which has 250 excerpts, and the 4 few-example classes). There is a new rhythm class 'Foxtrot'.

There are also 4 other classes (PasoDoble, Salsa, SlowWaltz, WcSwing) with relatively low number of tracks (47 on average). Using them lead to a really unbalanced dataset, but class imbalance could an interesting issue to have to deal with, so we publish them anyway.

Rhythm class	Ballroom	Extended Ballroom v1
Chacha	111	455
Jive	60	350
Quickstep	82	497
Rumba	98	470
Samba	86	468
Tango	86	464
Viennesewaltz	65	252
Waltz	110	529
Foxtrot		507
Pasodoble		53
Salsa		47
Slowwaltz		65
Wcswing		23
Total	**698**	**4180**

Table 1. Rhythm-class distribution for the Ballroom and the Extended Ballroom datasets.

It can be noted here that the Ballroom Dataset is not strictly included in the new Extended Ballroom: some albums and tracks originally present in 2004 are not sold anymore by the website. Only 343 tracks of the Ballroom Dataset are included in the Extended Ballroom Dataset (the intersection of the datasets was done with the ids and album/title of the tracks, but not with the content). Merging the two datasets does not make sense as the audio qualities are really different.

2.2 Annotations (B)

For each of the 4.180 tracks of the Extended Ballroom dataset we provide the following annotations: tempo (in bpm), rhythm class (or genre), artist, song title and album name. We also indicate for each track, its links to other tracks of the dataset based on their similarity. We define below four similarity categories.

2.2.1 Origin of the anotations (B1)

The rhythm class (or genre), the tempo (in bpm), the artist, the title and the album name were all extracted automatically from the source website. The similarity annotations were done in a semi-automatic way. Possible duplicates were found using the audio-fingerprint algorithm AudioPrint [4], and the similarity between title names (edit distance). All of candidates were then checked manually.

2.2.2 Concept definitions (B21)

We do not provide the definition of tempo (in bpm) and rhythm class (or genre) since they were extracted from the website. Since the website provides music for dancers learning ballroom dances, the definition is based on usage (no specific definition is given on the website).

The potential similarity between two tracks is defined using four categories (inspired by the work of [5]):

Exact repetition: the time-frequency domains of the two tracks are highly similar (different audio encoding, or slight temporal delay).

Time repetition: same track, beginning at two different instants.

Karaoke repetition: two identical tracks, but with the singing voice not present in one of them, or replaced by an instrument.

Version repetition: the two excerpts are the same, but played in a different way (it can be studio/live versions, or the same excerpts transposed, or a same song played with different instruments).

The distribution of these repetition categories is indicated in Table 2.

Exact	Time	Karaoke	Version
248	16	12	257

Table 2. Distribution of the different types of repetitions in the Extended Ballroom test-set.

2.2.3 Reliability (B32)

The tempo, the genre, the artist name and album name are all extracted from the website. We don't have any clues about how were made the annotations. However these annotations (tempo and genre) have been used for more than ten years without strong criticisms. In addition, no problems were found in the genre annotations on all the excerpts we listened to while creating/testing the database and looking for duplicates.

The reliability of the audio content of this database is one of the main challenge, as we can't distribute the raw audio along the metadata (for storage capacity and copyright reasons). A Python script is provided. With this script, everybody can download and more importantly check that all the downloaded tracks match exactly those we used. We provide a list of MD5 hashes of all the tracks we used in this dataset.

Finally, as the similarity annotations were done in a semi-automatic way (tracks with a similar title and a similar audio fingerprint were checked manually for duplicates), some duplicates/ repetitions may have been forgotten. However they should not represent more than a percent of the dataset as we tuned our automatic detection system to output a lot of false-positives.

2.2.4 Annotation tool (B34)

We refer the reader to the publication on Ramona & Peeters [4] for a complete description of the audio fingerprinting system.

2.3 Documentation (C)

2.3.1 Identification of the corpus (C1)

The identifier of this dataset is Extended Ballroom v1.

2.3.2 Storage (C2)

The Extended Ballroom dataset is can be downloaded at *http://anasynth.ircam.fr/home/media/ExtendedBallroom*. It is distributed as

- A single XML file which contains all meta-data and for each audio track its MD5 hash,
- A Python script that can download all corresponding audio tracks and check that MD5 hash correspondence,
- A readme file that describes the XML format.

3. APPLICATIONS

This dataset has similar applications as the Ballroom dataset. It is useful for genre/rhythm-class recognition systems as well as tempo estimation algorithms.

An important contribution of this paper is to provide a list of duplicated tracks, version repetitions, karaoke repetitions. One of the main interest of this dataset is the number of different versions that can be found for some songs, each version having a different instrumentation, singers, tonality, tempo. These annotations could be useful for building cover song detection systems.

In this new Extended Ballroom dataset, around 4.000 tracks (annotated in rhythm class) are available. It is around 6 times the previous Ballroom Dataset. Having this increased number of tracks can be really useful to methods based on Deep Neural Networks as they need a lot of input data. For reference, current state-of-the-art rhythm description system achieves 96.0% mean-over-classes recall on the Ballroom Dataset, and 94.9% on the Extended Ballroom [2].

4. REFERENCES

[1] 5th International Conference on Music Information Retrieval (ISMIR 2004) Rhythm description contest. http://mtg.upf.edu/ismir2004/contest/rhythmcontest/.

[2] Ugo Marchand and Geoffroy Peeters. Scale and shift invariant time/frequency representation using auditory statistics: application to rhythm description. In *IEEE International Workshop on Machine Learning for Signal Processing*, September 2016.

[3] Geoffroy Peeters and Karen Fort. Towards a (better) definition of the description of annotated m.i.r. corpora. October 2012.

[4] Mathieu Ramona and Geoffroy Peeters. Audioprint: An efficient audio fingerprint system based on a novel cost-less synchronization scheme. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 818–822. IEEE, 2013.

[5] Bob L Sturm. The gtzan dataset: Its contents, faults, and their effects on music genre recognition evaluation. *IEEE Transactions on Audio, Speech and Language Processing*, 2013.