DA-TACOS: A DATASET FOR COVER SONG IDENTIFICATION AND UNDERSTANDING

Furkan Yesiler[1] **Chris Tralie**[2] **Albin Correya**[1] **Diego F. Silva**[3]
Philip Tovstogan[1] **Emilia Gómez**[1 4] **Xavier Serra**[1]
[1] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
[2] Department of Mathematics And Computer Science, Ursinus College, USA
[3] Departamento de Computação, Universidade Federal de São Carlos, Brazil
[4] Joint Research Centre, European Commission, Seville, Spain
furkan.yesiler@upf.edu, ctralie@alumni.princeton.edu

ABSTRACT

This paper focuses on Cover Song Identification (CSI), an important research challenge in content-based Music Information Retrieval (MIR). Although the task itself is interesting and challenging for both academia and industry scenarios, there are a number of limitations for the advancement of current approaches. We specifically address two of them in the present study. First, the number of publicly available datasets for this task is limited, and there is no publicly available benchmark set that is widely used among researchers for comparative algorithm evaluation. Second, most of the algorithms are not publicly shared and reproducible, limiting the comparison of approaches. To overcome these limitations we propose Da-TACOS, a DaTAset for COver Song Identification and Understanding, and two frameworks for feature extraction and benchmarking to facilitate reproducibility. Da-TACOS contains 25K songs represented by unique editorial metadata plus 9 low- and mid-level features pre-computed with open source libraries, and is divided into two subsets. The Cover Analysis subset contains audio features (e.g. key, tempo) that can serve to study how musical characteristics vary for cover songs. The Benchmark subset contains the set of features that have been frequently used in CSI research, e.g. chroma, MFCC, beat onsets etc. Moreover, we provide initial benchmarking results of a selected number of state-of-the-art CSI algorithms using our dataset, and for reproducibility, we share a GitHub repository containing the feature extraction and benchmarking frameworks.

1. INTRODUCTION

Cover songs play an important role in the history of recorded music. Weinstein [42] argues that cover songs are peculiar to rock music, and some iconic early rock bands, like The Beatles, The Rolling Stones and Led Zeppelin, recorded cover songs at the beginning of their careers. Artists from other genres eventually followed this trend of reinterpreting recorded musical works. More recently, audio and video online streaming platforms have given rise to a great volume of fan versions of numerous original songs, including so-called "Youtube covers". Cataloguing and tracking cover versions of songs are important both from a historical and a legal standpoint, since there is sometimes a fine line between creative license and plagiarism [22]. However, this task often requires automation via content-based MIR strategies due to the explosion of recordings across many repositories.

Automatic Cover Song Identification (CSI) systems must contend with the myriad changes of musical facets that can occur among versions. While cover songs may share some musical characteristics, such as melody, harmony or chord progression, they are not identical musical works. According to Serrà [28], one can categorize musical transformations between cover versions into 8 main groups: timbre (due to production techniques and/or due to instrumentation), tempo, timing, structure, key, harmonization, lyrics and noise. Given this, the vast majority of CSI systems focus solely on the tonal content [3, 8, 31, 33, 35], a characteristic thought to be least altered between a song and its cover versions. Such systems work on top of features which are invariant to these transformations, incorporating techniques such as beat-synchronous features [12] to control for changes in tempo, or Optimal Transposition Index (OTI) [29] to control for changes in key.

In spite of Serrà's taxonomy and intuition about what makes a cover, to our knowledge, there are no large-scale studies quantifying the extent to which the aforementioned musical attributes change among cover versions. Furthermore, a variety of CSI algorithms have been designed under different assumptions about what makes a cover, each with different goals and trade-offs in mind, but the community lacks a large-scale open source dataset to compare their performance; the largest benchmark set to date is the SecondHandSongs dataset (SHS), a subset of the Million

Song Dataset [4], with 18,196 songs but this contains proprietary features. Motivated by both of these problems, we propose a new dataset, which we call "Da-TACOS"; a DaTAset for COver Song Identification and Understanding (Section 3). This dataset consists of 25,000 songs with a variety of audio descriptors, including low level features such as frame level HPCPs/MFCCs, beat onsets, and higher level information such as key, tempo, and audio tags. We then split our collection of songs into two subsets. The *Cover Analysis subset* is used to quantify what makes a cover by looking at changes in key, local onset deviation, tempo, rhythm, and audio tag descriptions (Section 4). In our analyses, we also introduce some tools not previously seen in the MIR community, such as ShapeDNA [25] and topological time series analysis [26]. The *Benchmark subset* is used for detailed benchmarking of a set of representative CSI algorithms selected across years of work on this topic (Section 5). After we set the stage with our preliminary experiments, we expect this dataset will enable researchers to continue to explore both the *what is a cover* question and *benchmarking* in more detail.

2. RELATED WORK

Most CSI systems have 3 main building blocks [23]: feature extraction, feature post-processing, and similarity estimation. An extensive review of the traditional cover song identification systems can be found in Serrà et al. [30]. In this section, we present a brief overview of these building blocks, the datasets for this task, and the observed limitations of current approaches.

Traditional CSI systems begin with low level feature extraction from the audio. The most common audio descriptors used in those systems are Pitch Class Profiles (PCP), or Chroma features, which represent the tonal content of songs via the octave-folded energies for each of the 12 pitch classes used in Western music theory. The task at hand informs which strategy is chosen, and to improve robustness, different variants of Chroma features [31, 34] were used in the CSI literature for this task. Moreover, many audio descriptors such as pitch salience [27], chord profiles [16], self-similarity MFCC [39], cognition-inspired descriptors [2] were also utilized; however, they suffer from lower performance scores in isolation compared to PCPs.

After the feature extraction, several feature post-processing steps can be applied to achieve invariances in several musical facets such as key, tempo and structure. Key invariance can be obtained using OTI [29] or the 2D Fourier Transform Magnitude (2DFTM) coefficients of the tonal features [14]. Beat-synchronous features are used to achieve tempo invariances [12, 38]. Similarity Matrix Profile (SiMPle) [35], which is a "representation of the similarity join between subsequences", can be useful to control for structural invariance or to obtain audio thumbnails of songs which can be later used to estimate the similarity of two songs [34].

The final step of this general CSI system framework is the similarity estimation. For certain representations, e.g. 2DFTM, this step may consist of only a simple distance function such as Euclidean or Cosine distances. However, for more accurate smaller scale algorithms, a quadratic alignment algorithm is often used to obtain tempo or structural invariance. Since global alignments between versions don't often exist in practice, CSI researchers put more emphasis on the Local Alignment methods such as Smith-Waterman algorithm [36] that is designed to detect alignments among all possible subsequences by incorporating local constraints. Depending on these constraints, many versions of this algorithm were proposed, e.g. *Qmax* [31] and *Dmax* [8]. The longest alignment is taken as the cover similarity measure/distance generally after normalizing it to the length of the reference track.

A number of previous works also explored combining different features and similarity measures to improve their systems. Salamon et al. [27] combine the distance values obtained with *Qmax* for melody, accompaniment and bass line. Chen et al. [8] use a technique called Similarity Network Fusion (SNF) [41] to integrate the similarity matrices obtained with *Qmax* and *Dmax* for the final similarity estimation. Tralie [38] uses SNF to combine cross-similarity matrices obtained with using HPCP and MFCC features to get a final similarity score.

Over the years, new methods were proposed by the MIR community to solve specific problems of the CSI task; however, due to the O(N^2) complexity of local alignment algorithms, some have focused on alternative algorithms that scale better. With the introduction of the SHS dataset [4], techniques such as audio fingerprinting [3], database pruning strategies [24] and multi-modal approaches [10] were also explored in the literature. But the performance scores of these scalable approaches obtained for SHS were not satisfactory. Thus, a trade-off between efficiency and robustness exists in the CSI task as well as in many other MIR tasks, and the "Holy Grail" algorithm for CSI that is both scalable and robust is still missing.

Although a large amount of previous works exist for CSI task, there are only a few public datasets available for benchmarking. Covers80, released by Ellis [13], contains 80 cliques, or cover groups, with 2 songs per clique. Although small in terms of size, this dataset includes audio files of the songs, which provides an opportunity for developing new features or fine-tuning the existing feature extraction algorithms. The YoutubeCovers dataset [33] contains 50 cliques with 7 songs per clique, and instead of audio files, pre-computed Chroma, CENS and Chroma DCT-Reduced log Pitch (CRP) features are included in this dataset. SHS is a subset of Million Song Dataset, and it contains pre-computed features extracted with EchoNest API [1] for 12960 songs in 4128 cliques for the training subset and 5236 songs in 726 cliques for the test subset [4]. Although comparatively larger in size, SHS comes with features pre-computed with proprietary algorithms which makes it impossible to reproduce or even use other audio descriptors for the CSI task.

Based on the limitations of current CSI systems and dif-

[1] http://the.echonest.com/

ficulty in comparing them, we propose a new dataset, and public frameworks for feature extraction and benchmarking to give CSI research a uniform direction. Our contributions can be summarized as follows:

- The largest benchmark set with 15,000 songs including state-of-the-art audio features for CSI
- The Cover Analysis subset with 10,000 songs for musicological studies
- First large-scale quantitative analysis on modified musical characteristics
- Open Source frameworks for feature extraction and benchmarking specifically created for the CSI task
- Open Source implementations of seven state-of-the-art systems and their initial benchmarking results

3. DA-TACOS: DATASET FOR COVER SONG IDENTIFICATION

For facilitating benchmarking and providing a set of analyses regarding links among cover songs, here, we propose a new dataset for CSI research. Da-TACOS, a DaTAset for COver Song Identification and Understanding, contains commercial or live recordings of 25,000 songs that are distributed into 2 subsets: the *Cover Analysis subset* and the *Benchmark subset* with 10,000 and 15,000 songs, respectively. The song annotations are collected from SecondHandSongs.com [2], and are licenced under Creative Commons BY-NC 3.0 [3]. Metadata for each song includes song title, name of the performer, original song title, name of the original writer and release year.

We have also matched the original metadata with MusicBrainz [4] [37] to obtain the MusicBrainz ID (MBID), length and genre tags. Most songs belong to rock, pop, metal and jazz genres. The average length of songs in the dataset is 3.59 minutes.

Along with the metadata, we share low- and mid-level features pre-computed with open source feature extraction libraries from MIR community; a comprehensive list can be found in Table 1. To increase the reliability of our results and assist future works, we share a common feature extraction framework, with which we obtained the feature values, in our GitHub repository.

In particular, Da-TACOS addresses two needs of the current state of CSI research. First, in Section 2, we mentioned the difficulty of benchmarking CSI systems, and with this dataset, we take a step toward tackling this challenge. We provide a large set of pre-computed features that have been constantly used in CSI research, and we provide initial benchmarking results of a selected number of state-of-the-art CSI systems. Second, to our knowledge, our benchmark subset is the largest dataset to date for comparing the performances of CSI systems. We see this as an opportunity to scale up CSI research to discover methods that are more likely to be used in real world scenarios. We believe that having a large dataset for benchmarking will have a positive effect on the direction of future research for this task.

[2] http://secondhandsongs.com
[3] https://creativecommons.org/licenses/by-nc/3.0/
[4] https://musicbrainz.org

	Benchmark	Cover Analysis	
HPCP	✓	✓	Essentia [7]
MFCC	✓	✓	""
Key	✓	✓	""
CENS	✓	✓	Librosa [21]
Tempogram	✓	✓	""
Beat Onsets	✓	✓	Madmom [5]
Tempo	✓	✓	""
CREMA	✓	✓	CREMA [18]
Auto-tagger		✓	Choi et. al [9]

Table 1. List of features provided in each subsets of Da-TACOS and the related feature extraction libraries used.

3.1 The Cover Analysis subset

Our first subset is dedicated to a series of analyses to understand the changes in musical characteristics when a new version of a song is created. This subset includes 10,000 songs in 5,000 cliques, a pair of cover songs for each clique. Out of all songs, we were able to match 6,821 songs with a MBID. The information regarding feature extraction and results of our analyses can be found in Section 4.

3.2 The Benchmark subset

The second subset of Da-TACOS is designed for benchmarking purposes. This subset includes total of 15,000 songs: 13,000 songs in 1000 cliques with 13 songs each, and 2,000 songs that do not belong to any clique, acting as noise in the data. For this subset 10,027 songs have MBIDs, and an initial benchmark of a selected set of CSI systems can be found in Section 5.

4. WHAT IS A COVER?

In this section, we exploit the features to explore the frequency and intensity of a subset of Serrà's [28] posited changes between cover versions. The analyses below are performed on the Cover Analysis subset of Da-TACOS. While key and tempo are straightforward to compare, we devise custom distance measures to compare timing, structure, and semantic aspects, e.g. instrumentation, genre. In these latter cases, we compare distributions of the corresponding distances between true cover pairs in this subset to all other non-cover pairs in the subset. To quantify the extent to which the true cover and non-cover distributions differ in these cases, we report the 2-sample Kolmogorov-Smirinov (KS) score, with its associated p-value, which indicates the statistical significance of the difference between two distributions.

4.1 Results

4.1.1 Key

Using a key estimation algorithm [15], we considered all pairs with both songs exceeding a confidence of 0.75,

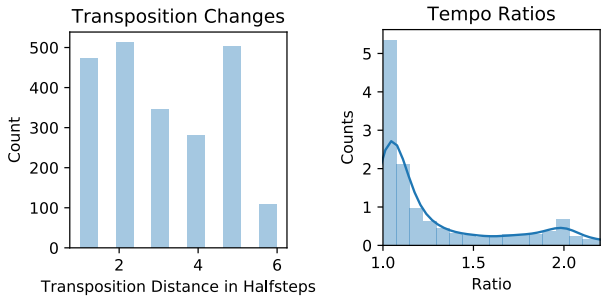

Figure 1. (Left) Distribution of halfsteps between key estimates for cover pairs with a reported key change. (Right) Distribution of tempo ratios between cover pairs.

which was 4288/5000 pairs. Among these, 69.3%, were reportedly in a different key. The distribution of said halfstep shifts is shown in Figure 1. Thus, the use of OTI in many CSI algorithms is justified. One caveat is that the key estimation algorithms report a single estimate which is either major or minor. Under this scheme, the key estimation algorithm reported that 17.5% of the pairs shifted from major to minor. However, upon spot checking, it was clear that many of these examples either switched keys at different times, or they were in modes beyond major and minor. In the absence of more sophisticated algorithms, expert ear trained individuals would be needed to determine how often changes beyond simple transpositions occur.

4.1.2 Tempo

We now examine tempo ratios between cover pairs by picking out the tempo with the maximum confidence from a state-of-the-art tempo estimator [6]. The right of Figure 1 shows the results. There is a slight peak around 2 which is likely due to "octave errors" from pieces which can be subdivided into 4/4. Beyond that, at the first quartile is a 1.03x change in tempo, in the second quartile is a 1.11x change in tempo, and in the third quartile is a 1.53x change in tempo. Thus, half of the songs are quite stable, but in the 50-75% quartile, we have a fair number of songs with a significant tempo change which can't easily be explained by a direct tempo doubling mistake, and which are likely "real." Hence, tempo is often a factor which needs to be controlled for when analyzing cover versions.

4.1.3 Structure

One particularly successful approach to multiscale music structure analysis uses eigenvectors of the Graph Laplacian, or "spectral clustering" [19]. While one can compare agreement of this technique to that of human annotators on the same piece of music [18], this representation does not immediately extend across versions of songs. We instead use the eigen*values* of the Graph Laplacian, which we stack up into a Euclidean vector which can be compared across songs. This has been referred to as "Shape DNA" in the context of 3D shape analysis of triangle meshes [25]. In our case, we use feature fused SSMs [40] downsampled to a common dimension of 256×256, followed by 30 eigenvalues of a random walk Laplacian.

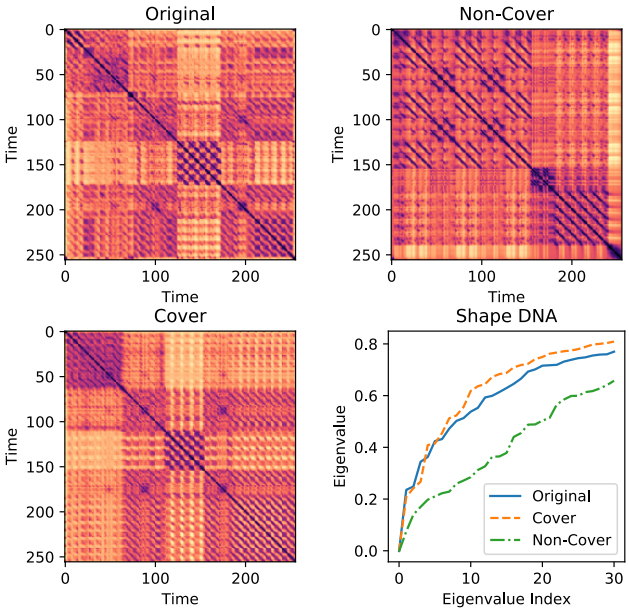

Figure 2. An example of fused similarity matrices of "The Wizard" by Uriah Heep (upper left), a cover by Blind Guardian (lower left), and "Million Pieces" by The Piano Tribute Players (upper right), which is unrelated. The corresponding Shape DNAs are shown in the lower right.

Figure 2 shows an example of Shape DNA between a pair of covers and a third, unrelated song. Even though the cover pairs' similarity matrices do not align perfectly and contain other variations, their shape DNAs are close, while they are both different from an unrelated song with a different structure. Figure 3 shows the distributions of shape DNA differences between true cover and non-cover pairs. The KS score between the two distributions is 0.22 ($p \ll 0.001$), indicating that while large structural changes do occur between cover versions (e.g. added/deleted sections), it is overall more likely for cover songs to share structure than random pairs of songs.

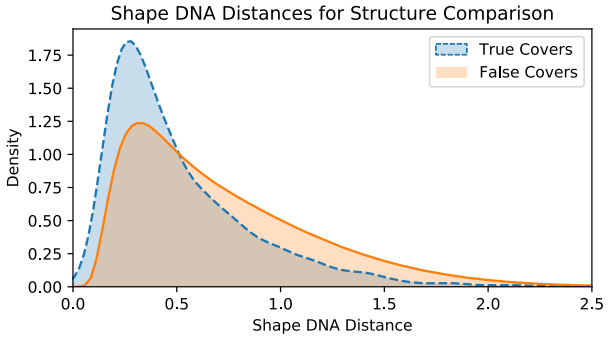

Figure 3. Distributions of shape DNA differences between pairs of songs as a means of assessing structural changes.

4.1.4 Timing

We now turn to timing, which we define as local changes in tempo over time. We first extract N beat onset estima-

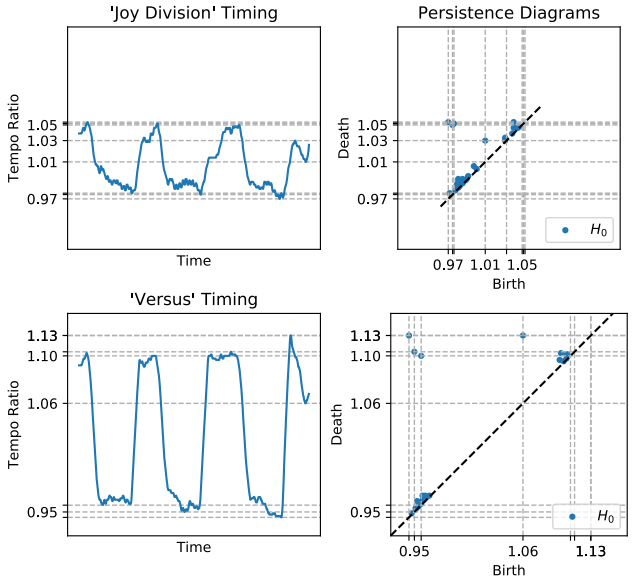

Figure 4. An example of $r[t]$ functions and their associated persistence diagrams for the song "24 Hours" by Joy Division and Versus. Both songs speed up in the chorus and slow down in the verse, so they each contain several local mins with high persistence which are born during the verses. They each also contain some low amplitude wobbling which shows up as dots near the diagonal.

tion times $b[t], t = 1, 2, \ldots N$ using the technique of Krebs [17], down to a resolution of 10 milliseconds. We then extract unit-less local tempo estimates by convolving $b[t]$ with a Gaussian derivative $b'[t] = b[t] * (-te^{-t^2/2})$, followed by a sliding window average of width 20 to smooth out noise. Finally, we divide $b'[t]$ by its median to obtain a relative, tempo-normalized local tempo deviation $r[t]$; $r[t] > 1$ if a song has sped up locally, and $r[t] < 1$ if it has slowed down locally.

The left column of Figure 4 shows $r[t]$ for two different versions of the same song. Note the multi-scale features of $r[t]$, from small wobbles in tempo to large changes that persist over a section. To capture all scales in one distance measure which can tolerate missing beats and added/deleted sections, we turn to the "lower star filtration," a watershed method from topological data analysis [11]. This summarizes a time series in a "persistence diagram"[5] (PD). This has been used, for instance, on speed time series of drivers to quantify driving behavior [26].

The right column of Figure 4 shows PDs for the $r[t]$ functions for an example cover pair, with the birth and death values of the points with 4 largest "persistence" (death-birth) marked. To compare PDs between two different songs, we use persistence images [1], which transform a diagram into birth-persistence space and place a Gaussian over each point, whose magnitude is proportional to the persistence. Figure 5 shows the distributions of Euclidean distances between peristence images for true cover

[5] A multiset of points whose x-coordinates correspond to local mins where pools of water form as water rises from bottom to top ("birth events"), and and whose y-coordinates correspond to local maxes paired to these mins where two pools merge together ("death events")

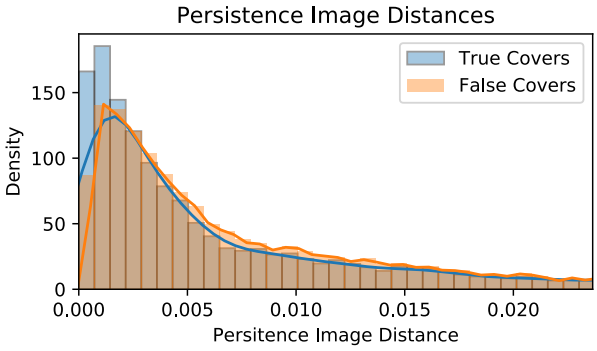

Figure 5. Distribution of persistence image distances of lower star filtrations relative tempo functions between pairs of songs.

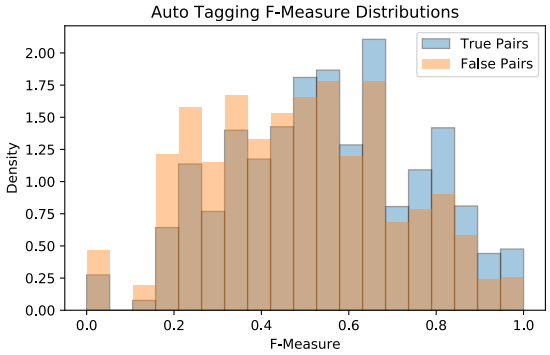

Figure 6. Distributions of f-measures for auto tagging for true cover pairs and non-cover pairs.

pairs and non-cover pairs. Though the distributions are quite similar, the KS score is 0.095 ($p \ll 0.001$), indicating that though relative timing can be different (as evidenced by Figure 4 where one song speeds up more in the chorus relative to the other), the difference is less for covers than for random pairs.

4.1.5 Semantic Aspects

To analyze semantic aspects of the songs such as mood, instrumentation and "genre" without explicitly defining them, we turn to auto tagging techniques of Choi et al. [9] which use log-mel spectrograms as input to return a set of tags which qualitatively describe the songs. Since the auto tagger returns many tags with low confidence, we only take tags which are in the 90$^{\text{th}}$ percentile over all confidences, which is a confidence value of 0.062. If p is the fraction of tags in song A contained in the set of tags for song B, and r is the fraction of tags in song B contained in the tags for song A, then the *f-measure* between two songs is defined as $2pr/(p + r)$, which is 1 if they are in complete agreement and 0 if they have nothing in common. Figure 6 shows the distribution of f-measures between true cover and non-cover pairs. While the distributions are overall quite similar, the f-measures are skewed slightly lower for non covers. The KS score between the two distributions is 0.118 ($p \ll 0.001$), indicating these two distributions are

different by more than chance; thus, we can conclude that although less frequent than between two random songs, stylistic changes occur between cover pairs.

5. BENCHMARKING

As mentioned in Section 3, Da-TACOS contains a benchmark subset of 15,000 songs for comparative algorithm evaluation. In this section, we present the results on seven different state-of-the-art algorithms on this data. To the best of our knowledge, this is the first work comparing these algorithms for CSI on a publicly available, large-scale dataset with features obtained with open source algorithms.

5.1 Methodology

One of the main limitations of current CSI research is the lack of a public framework to compare the performance of different systems. We acknowledge that the audio cover song identification task in the Music Information Retrieval Evaluation eXchange (MIREX) [6] addresses this. However, MIREX data is not publicly available, and it restricts the evaluation to a limited time window per year. According to the results from previous MIREX, [32] is still an algorithm which may be considered in the state-of-the-art CSI system. We chose to benchmark six other unsupervised algorithms which more recently presented good results on CSI for comparison in this competition [8, 14, 31, 35, 38, 39]. In their original implementations, these algorithms differ with the features used, a large scale or small scale design goal, their ability to combine distance measures or fusing more than one feature set [8, 38], exploitation of network structure of songs [8, 38], the application of beat-synchronous features [14, 38, 39], or a combination of these properties. In our work, we have the opportunity to control for implementation details that can greatly impact performance [20] both by sharing features across all algorithms, and by using common implementations of some sub-algorithms, including OTI, Similarity Network Fusion (SNF) [41] (for [38] and [8]), and QMax alignment [31].

5.2 Results

The empirical evaluation of CSI algorithms is another point in which published papers differ greatly. Commonly, different authors use different subsets of evaluation measures. For this reason, we used a large number of evaluation measures assessing the results, namely Mean Rank (MR), Mean Reciprocal Rank (MRR), Median Rank (MDR), Mean Average Precision (MAP), and the counting of correctly identified versions in top 1 and top 10.

In addition to using HPCPs as Chroma features for all the algorithms, we also use CREMA chord model features [18], sampled at the same rate, as a drop-in replacement for Chroma on all algorithms.

Table 2 presents the results obtained by all the algorithms considered in our evaluation, with a simplified version of Tralie's early fusion [38] which uses a weighted average in the early fusion stage instead of SNF for speed.

		MR	MRR	MDR	MAP	Top 1	Top 10
FTM2D [14]	H	207	0.314	15	0.126	3954	6131
	C	155	0.523	1	0.275	7185	9072
Simple [35]	H	358	0.362	13	0.165	4916	6361
	C	142	0.555	1	0.332	7739	9391
Dmax [8]	H	155	0.562	1	0.292	7939	9320
	C	134	0.571	1	0.322	7981	9611
LateFusion [8]	H	210	0.604	1	0.410	8761	9880
	C	177	0.621	1	**0.454**	8897	10223
Qmax [31]	H	119	0.606	1	0.333	8630	9931
	C	113	0.611	1	0.365	8625	10212
Qmax* [32]	C	**104**	0.619	1	0.373	8766	10246
SSM [39]	M	434	0.273	39	0.096	3540	5139
EarlyFusion [38]	H	116	**0.680**	1	0.426	**9843**	**10861**
	C	120	0.672	1	0.416	9667	10829

Table 2. Performance statistics of all algorithms. H stands for HPCP, C for CREMA and M for MFCC.

Overall, we find the CREMA improves results over HPCP, which suggests an adoption of CREMA for future research. We are particularly surprised at how well the large-scale FTM2D algorithm performs with CREMA.

6. CONCLUSION

In this work, we have presented a new public dataset, Da-TACOS, for analyzing how a number of musical facets vary among cover songs and benchmarking CSI systems. Our "what is a cover" analysis takes Serrà's [28] categories of modifiable musical characteristics as a basis, and the results demonstrate large variations between cover pairs across all of the aspects we examined, which supports Serrà's claims on the subject. However, the same analyses among non-cover pairs show a larger variation than cover pairs, and this can be interpreted as there are some links remaining among cover songs.

Moreover, we created a framework that includes open source implementations of seven state-of-the-art unsupervised CSI algorithms to facilitate the future work in this line of research. Using this framework, researchers can easily compare existing algorithms on different datasets, and we encourage all CSI researchers to incorporate their algorithms into this framework in order to support Open Science principles. Our feature extraction and benchmarking frameworks as well as instructions to can be found in our GitHub repository [7].

Our future work includes constructing several other subsets based on various characteristics of songs (e.g. subsets based on genre and release year), as well as training sets for supervised algorithms, to identify the further needs of CSI research. We believe that a better understanding of the relationships among cover songs is valuable both for musicological aspect of this line of research and for advancing the state of the art in CSI research.

[6] https://www.music-ir.org/mirex/wiki

[7] https://github.com/furkanyesiler/acoss

7. ACKNOWLEDGMENTS

This work is partially supported by the MIP-Frontiers project, the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 765068, and by TROMPA, the Horizon 2020 project 770376-2.

8. REFERENCES

[1] Henry Adams, Sofya Chepushtanova, Tegan Emerson, Eric Hanson, Michael Kirby, Francis Motta, Rachel Neville, Chris Peterson, Patrick Shipman, and Lori Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *The Journal of Machine Learning Research*, 18(1):218–252, 2017.

[2] Jan Van Balen, Dimitrios Bountouridis, Frans Wiering, and Remco Veltkamp. Cognition-inspired descriptors for scalable cover song retrieval. In *15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, Taipei, Taiwan, 2014.

[3] Thierry Bertin-Mahieux and Daniel P.W. Ellis. Large-scale cover song recognition using hashed chroma landmarks. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, USA, 2011.

[4] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In *12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, Miami, Florida, USA, 2011.

[5] Sebastian Böck, Filip Korzeniowski, Jan Schlüter, Florian Krebs, and Gerhard Widmer. Madmom: A new Python audio and music signal processing library. In *Proc. of the 24th ACM International Conference on Multimedia*, pages 1174–1178, Amsterdam, The Netherlands, 2016.

[6] Sebastian Böck, Florian Krebs, and Gerhard Widmer. Accurate tempo estimation based on recurrent neural networks and resonating comb filters. In *16th International Society for Music Information Retrieval Conference (ISMIR 2015)*, pages 625–631, Malaga, Spain, 2015.

[7] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, José Zapata, and Xavier Serra. Essentia: An audio analysis library for music information retrieval. In *14th Conference of the International Society for Music Information Retrieval (ISMIR 2013)*, Curitiba, Brazil, 2013.

[8] Ning Chen, Wei Li, and Haidong Xiao. Fusing similarity functions for cover song identification. *Multimedia Tools and Applications*, 77(2):2629–2652, 2018.

[9] Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. Convolutional recurrent neural networks for music classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP))*, pages 2392–2396, New Orleans, Louisiana, USA, 2017.

[10] Albin Correya, Romain Hennequin, and Mickaël Arcos. Large-scale cover song detection in digital music libraries using metadata, lyrics and audio features. *arXiv preprint arXiv:1808.10351*, 2018.

[11] Herbert Edelsbrunner and John Harer. *Computational topology: An introduction*. American Mathematical Soc., 2010.

[12] Daniel P.W. Ellis. Identifying "cover songs" with beat-synchronous chroma features. *MIREX 2006*, pages 1–4, 2006.

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

[14] Daniel P.W. Ellis and Thierry Bertin-Mahieux. Large-scale cover song recognition using the 2D Fourier Transform magnitude. In *13th International Society for Music Information Retrieval Conference (ISMIR 2012)*, Porto, Portugal, 2012.

[15] Emilia Gómez. Tonal description of polyphonic audio for music content processing. *INFORMS Journal on Computing*, 18(3):294–304, 2006.

[16] Maksim Khadkevich and Maurizio Omologo. Large-scale cover song identification using chord profiles. In *14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, pages 233–238, Curitiba, Brazil, 2013.

[17] Florian Krebs, Sebastian Böck, and Gerhard Widmer. An efficient state-space model for joint tempo and meter tracking. In *16th International Society for Music Information Retrieval Conference (ISMIR 2015)*, pages 72–78, Malaga, Spain, 2015.

[18] Brian McFee and Juan P. Bello. Structured training for large-vocabulary chord recognition. In *18th International Society for Music Information Retrieval Conference (ISMIR 2017)*, pages 188–194, Suzhou, China, 2017.

[19] Brian McFee and Daniel P.W. Ellis. Analyzing song structure with spectral clustering. In *15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, Taipei, Taiwan, 2014.

[20] Brian McFee, Jong Wook Kim, Mark Cartwright, Justin Salamon, Rachel M. Bittner, and Juan P. Bello. Open-source practices for music signal processing research: Recommendations for transparent, sustainable, and reproducible audio research. *IEEE Signal Processing Magazine*, 36(1):128–137, 2019.

[21] Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in Python. In *Proc. of the 14th Python in Science Conference*, pages 18–25, Austin, Texas, USA, 2015.

[22] Emily Miao and Nicole E. Grimm. The blurred lines of what constitutes copyright infringement of music: Robin Thicke v. Marvin Gaye's estate. *Westlaw Journal Intellectual Property*, 20:1, 2013.

[23] Julien Osmalskyj. *A Combining Approach to Cover Song Identification*. PhD thesis, University of Liege, Belgium, 2017.

[24] Julien Osmalskyj, Sébastien Piérard, Marc Van Droogenbroeck, and Jean-Jacques Embrechts. Efficient database pruning for large-scale cover song recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 714–718, Vancouver, Canada, 2013.

[25] Martin Reuter, Franz-Erich Wolter, and Niklas Peinecke. Laplace–Beltrami spectra as 'Shape-DNA' of surfaces and solids. *Computer-Aided Design*, 38(4):342–366, 2006.

[26] David Rouse, Adam Watkins, David Porter, John Harer, Paul Bendich, Nate Strawn, Elizabeth Munch, Jonathan DeSena, Jesse Clarke, Jeff Gilbert, Sang Chin, and Andrew Newman. Feature-aided multiple hypothesis tracking using topological and statistical behavior classifiers. In *Signal Processing, Sensor/Information Fusion, and Target Recognition XXIV*, page 94740L, 2015.

[27] Justin Salamon, Joan Serrà, and Emilia Gómez. Melody, bass line, and harmony representations for music version identification. In *Proc. of the 21st Int. World Wide Web Conf. (WWW 2012): 4th Int. Workshop on Advances in Music Information Research (AdMIRe 2012)*, pages 887–894, Lyon, France, 2012.

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

[29] Joan Serrà, Emilia Gómez, and Perfecto Herrera. Transposing chroma representations to a common key. In *IEEE CS Conference on The Use of Symbols to Represent Music and Multimedia Objects*, pages 45–48, 2008.

[30] Joan Serrà, Emilia Gómez, and Perfecto Herrera. Audio cover song identification and similarity: background, approaches, evaluation, and beyond. In *Advances in Music Information Retrieval*, pages 307–332. Springer, 2010.

[31] Joan Serrà, Xavier Serra, and Ralph G. Andrzejak. Cross recurrence quantification for cover song identification. *New Journal of Physics*, 11(9):093017, 2009.

[32] Joan Serrà, Massimiliano Zanin, Cyril Laurier, and Mohamed Sordo. Unsupervised detection of cover song sets: Accuracy improvement and original identification. In *10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, pages 225–230, Kobe, Japan, 2009.

[33] Diego Furtado Silva, Vinícius M. A. de Souza, and Gustavo E. A. P. A. Batista. Music shapelets for fast cover song recognition. In *16th International Society for Music Information Retrieval Conference (ISMIR 2015)*, pages 441–447, Malaga, Spain, 2015.

[34] Diego Furtado Silva, Felipe Vieira Falcão, and Nazareno Andrade. Summarizing and comparing music data and its application on cover song identification. In *19th International Society for Music Information Retrieval Conference (ISMIR 2018)*, pages 732–739, Paris, France, 2018.

[35] Diego Furtado Silva, Chin-Chia Michael Yeh, Gustavo E.A.P.A. Batista, and Eamonn J. Keogh. SiMPle: Assessing music similarity using subsequences joins. In *17th International Society for Music Information Retrieval Conference (ISMIR 2016)*, pages 23–29, New York City, New York, USA, 2016.

[36] Temple F. Smith and Michael S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981.

[37] Aaron Swartz. Musicbrainz: A semantic web service. *IEEE Intelligent Systems*, 17(1):76–77, 2002.

[38] Christopher J. Tralie. Early MFCC and HPCP fusion for robust cover song identification. In *18th International Society for Music Information Retrieval Conference (ISMIR 2017)*, Suzhou, China, 2017.

[39] Christopher J. Tralie and Paul Bendich. Cover song identification with timbral shape sequences. In *16th International Society for Music Information Retrieval Conference (ISMIR 2015)*, Malaga, Spain, 2015.

[40] Christopher J. Tralie and Brian McFee. Enhanced hierarchical music structure annotations via feature level similarity fusion. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019.

[41] Bo Wang, Aziz M. Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity Network Fusion for aggregating data types on a genomic scale. *Nature methods*, 11(3):333, 2014.

[42] Deena Weinstein. The history of rock's pasts through rock covers. In T. Swiss, J. Sloop, and A. Herman, editors, *Mapping the Beat: Popular Music and Contemporary Theory*, pages 137–151. Blackwell Publishing, Malden, MA, USA, 1998.