STRUCTURE-BASED AUDIO FINGERPRINTING FOR MUSIC RETRIEVAL

Peter Grosche[1,3], **Joan Serrà**[4], **Meinard Müller**[2,3], **and Josep Ll. Arcos**[4]
[1] Saarland University, [2] Bonn University, [3] MPI Informatik
[4] Artificial Intelligence Research Institute (IIIA-CSIC)
{pgrosche,meinard}@mpi-inf.mpg.de, {jserra,arcos}@iiia.csic.es

ABSTRACT

Content-based approaches to music retrieval are of great relevance as they do not require any kind of manually generated annotations. In this paper, we introduce the concept of *structure fingerprints*, which are compact descriptors of the musical structure of an audio recording. Given a recorded music performance, structure fingerprints facilitate the retrieval of other performances sharing the same underlying structure. Avoiding any explicit determination of musical structure, our fingerprints can be thought of as a probability density function derived from a self-similarity matrix. We show that the proposed fingerprints can be compared by using simple Euclidean distances without using any kind of complex warping operations required in previous approaches. Experiments on a collection of Chopin Mazurkas reveal that structure fingerprints facilitate robust and efficient content-based music retrieval. Furthermore, we give a musically informed discussion that also deepens the understanding of this popular Mazurka dataset.

1. INTRODUCTION

The rapidly growing corpus of digitally available audio material requires novel retrieval strategies for exploring large collections and discovering music. One outstanding instance of content-based music retrieval is *query-by-example*: Given a query in the form of an audio recording (or just a short fragment of it), the goal is to retrieve all documents from a music collection that are somehow similar or related to the query. In this context, the notion of similarity used to compare different audio recordings (or fragments) is of crucial importance and largely depends on the respective application. Typical similarity measures assess timbral, melodic, rhythmic, or harmonic properties [2].

A further key aspect of music is its structure. Indeed, the automatic extraction of structural information from music recordings constitutes a central research topic within the area of music information retrieval [10]. One goal of structure analysis is to split up a music recording into segments and to group these segments into musically meaningful categories, such as chorus or verse. The structure is a highly characteristic property for many musical styles. Folk songs and children songs, for example, typically exhibit a strophic form, where one tune is repeated over and over again with changing lyrics. Popular music typically consists of a number of repeating verses connected by a refrain. In classical music, the structure (or musical form) is often more complex and offers more variability.

Besides being characteristic for a certain musical style, the structure and, in particular, the relative duration of its elements is also a good descriptor for a specific piece of music—irrespective of specific realizations or performances. Furthermore, the structure is invariant to changes in instrumentation or key and therefore allows for identifying different performances of the same piece. So far, only a few approaches exist that exploit structural similarity to facilitate music retrieval [1, 4, 6, 7]. Typically, these approaches are based on *self-similarity matrices* (SSMs) which in general play an important role for analyzing musical structures [10]. For computing an SSM, an audio recording is first transformed into a sequence of feature vectors and then all elements of the sequence are compared in a pairwise fashion using a local similarity measure. Repeating patterns in the feature sequence appear as parallel paths in the SSM, see Figure 1a. Revealing structural properties, SSMs can in turn be used for analyzing structural similarities of performances. To this end, one requires a similarity measure that compares entire SSMs while being invariant to temporal variations. In [6, 7], the SSMs are compared using a similarity measure that is based on a two dimensional version of dynamic programming. The approach proposed by Bello [1] is also based on SSMs, but employs a *normalized compression distance* (NCD) to assess their similarity, without requiring any alignment operations. Originally proposed for comparing protein structures in bioinformatics, the NCD can be regarded as a measure of the *information distance* of two objects where the Kolmogorov complexity is approximated using a standard compression algorithm, see [1].

Inspired by the work of Bello, we describe in this paper a simple yet effective approach for measuring structural similarities of music recordings. As first contribution, we introduce the concept of *structure fingerprints* which are compact structural descriptors of music record-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.
© 2012 International Society for Music Information Retrieval.

ings. Analogous to [1, 6, 7], our fingerprints are also derived from self-similarity matrices while avoiding any explicit determination of structure. Specifically, we use a bivariate variant of a Parzen-Rosenblatt kernel density estimation method for representing a given SSM by a probability density function (pdf) [13]. This has the desired effect of smoothing out temporal variations in the performances. As a result, unlike previous approaches, we do not require any complex distance measure. Instead, recordings can be compared efficiently using, e. g., the Euclidean distance between fingerprints. As second contribution, we report on extensive experiments using a large collection of Chopin Mazurkas. In particular, we show that structure fingerprints facilitate content-based music retrieval solely based on structural information and exhibit a high degree of robustness against performance variations. This makes the presented approach particularly suited for supporting traditional retrieval systems that assess harmonic similarities [2, 5, 8, 12]. Finally, as third contribution, we provide a musically informed discussion of problematic pieces and recordings which also deepens the understanding of the Mazurka dataset.

The remainder of this paper is organized as follows. In Section 2, we introduce our approach to computing structure fingerprints. Then, in Section 3, we describe our retrieval experiment and give a quantitative as well as musically informed discussion of the results. Conclusions are given in Section 4.

2. STRUCTURE FINGERPRINTS

In this section, we introduce our strategy for computing structure fingerprints that capture characteristics of a musical piece and, at the same time, are invariant to properties of a specific performance. We first introduce the underlying feature representation (Section 2.1) and the SSM variant (Section 2.2). In particular, we introduce various enhancement strategies that absorb a large degree of temporal and spectral variations. Then, in Section 2.3 we explain in detail how the fingerprints are derived from the SSMs.

2.1 Feature Representation

We first convert a given music recording into a sequence of chroma features, which have turned out to be a powerful mid-level representation for relating harmony-based music [1, 2, 5, 8, 10, 12]. The term *chroma* refers to the elements of the set $\{\mathrm{C}, \mathrm{C}^\sharp, \mathrm{D}, \ldots, \mathrm{B}\}$ that consists of the twelve pitch classes as used in Western music notation. Representing the short-time energy content of the signal relative to the pitch classes, chroma features do not only account for the close octave relationship in harmony, but also introduce a high degree of robustness to variations in timbre and instrumentation [8]. Furthermore, normalizing the features makes them invariant to dynamic variations.

In our implementation, we use a variant of chroma features referred to as CENS [1] features [8]. As main advantage, CENS features involve an additional temporal smoothing and downsampling step which leads to an increased robustness of the features to local tempo changes [8]. This property is crucial for obtaining structure fingerprints that are invariant to local variations in the performances. In our implementation, the resulting feature representation has a resolution of 1 Hz (one feature per second), where each vector is obtained by averaging over 4 seconds of the audio.

[1] *Chroma Energy Normalized Statistics* features, provided by the Chroma Toolbox www.mpi-inf.mpg.de/resources/MIR/chromatoolbox

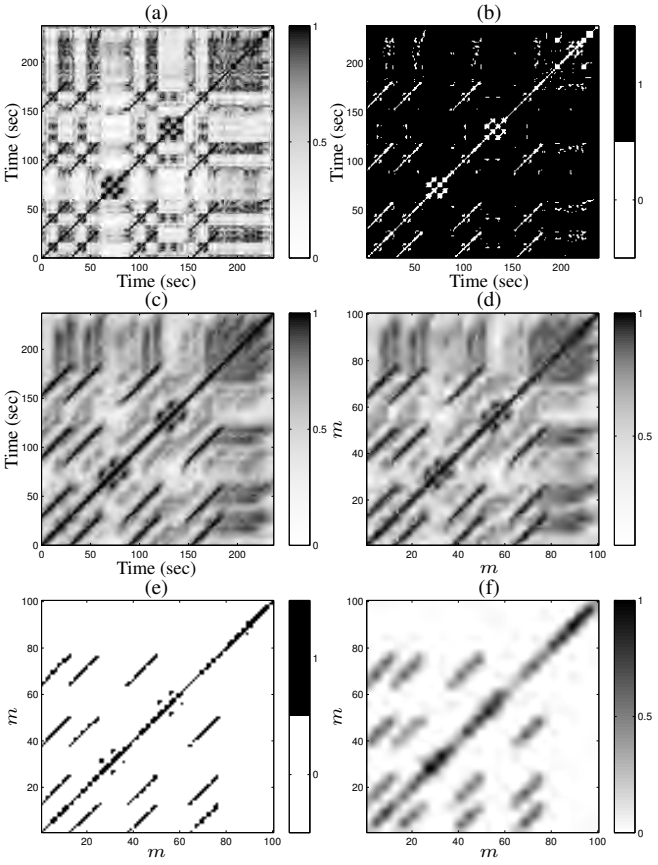

Figure 1: Computing structure fingerprints for an Ashkenazy (1981) performance of Chopin's Mazurka Op. 56 No. 1 with the musical form $A_1A_2BA_3CA_4D$. **(a)** SSM computed from CENS features. **(b)** Thresholded variant of (a) ($\kappa = 10$). **(c)** Path-structure enhanced SSM ($L = 12$). **(d)** Resampled SSM $\mathbf{S}_M^{\mathrm{fix}}$ ($M = 100$). **(e)** Thresholded variant of (d) ($\kappa = 10$). **(f)** Structure fingerprints (pdf estimated from (e), $\ell = 10$).

2.2 Self-Similarity Matrix

Let $X := (x_1, x_2, \ldots, x_N)$ be the feature sequence consisting of N normalized CENS features. Furthermore, let $\mathbf{s}$ be a similarity measure that allows for comparing two CENS vectors. In the following, we use the inner product between the normalized CENS vectors (cosine measure, which yields similarity values between 0 and 1). Then, a *self-similarity matrix* (SSM) is obtained by comparing all elements of X in a pairwise fashion [10]:

$$\mathbf{S}(n, m) := \mathbf{s}(x_n, x_m)$$

for $n, m \in [1 : N] := \{1, 2, \ldots, N\}$.

Figure 1a shows the resulting SSM for an Ashkenazy (1981) performance of Chopin's Mazurka Op. 56 No. 1

having the musical form $A_1A_2BA_3CA_4D$. The SSM reveals the repetitive structure (four repeating A-parts) in the form of diagonal paths of high similarity (dark colors).

2.2.1 Path-Structure Enhancement

Musical variations often lead to fragmented path structures of $\mathbf{S}$. To alleviate this problem, various matrix enhancement strategies have been proposed [1, 9, 12] with the idea to apply a smoothing filter along the direction of the main diagonal. This results in an emphasis of diagonal information and a denoising of other structures, see Figure 1c. In the presence of significant tempo differences, however, simply smoothing along the main diagonal may smear out important structural information. To avoid this, we use a strategy that filters the SSM along multiple gradients as proposed in [9]. In our experiments, we compute a simple moving average in windows corresponding to L seconds of audio and use five gradients covering tempo variations of -30 to $+30$ %. In the following, the enhanced SSM is again denoted as $\mathbf{S}$.

2.2.2 Resampling

A high degree of local tempo differences is already absorbed by the smoothing of the CENS features and the path-structure enhancement. Global differences in tempo of different performances of a piece of music, however, lead to SSMs that have different sizes. For deriving structure fingerprints that are invariant to such tempo differences, we apply the idea of [5] and introduce a simple resampling step that converts the $N \times N$ similarity matrix $\mathbf{S}$ into an $M \times M$ similarity matrix $\mathbf{S}^{\text{fix}}_M$, with M fixed to a suitable value:

$$\mathbf{S}^{\text{fix}}_M(n,m) := \mathbf{S}(\lfloor n\tfrac{N}{M} \rceil, \lfloor m\tfrac{N}{M} \rceil)$$

for $m, n \in [1 : M]$, where $\lfloor \cdot \rceil$ denotes rounding to the nearest integer.[2] Figure 1d shows an example for $\mathbf{S}^{\text{fix}}_M$.

2.2.3 Thresholding

We finally process the SSMs by suppressing all values that fall below a threshold. Analogous to [1, 12], we choose the threshold in a relative fashion by keeping κ % of the cells having the highest score. The motivation for this thresholding step is that only a certain amount of the cells of the SSM are expected to encode relevant structural information. The thresholding can then be regarded as some kind of denoising, where only relevant paths are retained, see Figure 1e. In the following, the resulting thresholded, resampled, and path-structure enhanced SSM is denoted as $\hat{\mathbf{S}}^{\text{fix}}_M$. Figure 1e also emphasizes the importance of the path-structure enhancement, as directly applying the thresholding operation on the original SSM does not lead to the desired denoising effect, see Figure 1b.

2.3 Probability Density Estimation

The four repeating A-parts of our Mazurka example are clearly revealed by $\hat{\mathbf{S}}^{\text{fix}}_M$ in the form of diagonal paths, see Figure 1e. However, as the structural information is contained in only a few cells of the thresholded SSM (in other words, the resulting matrix is sparse), small temporal variations in performances may lead to large distances when directly comparing these matrices in a pointwise fashion. As a result, some kind of tolerance to temporal variations is required in the similarity measure, as e. g., introduced by the similarity measures based on dynamic programming used in [6, 7] and the NCD used by Bello in [1].

Avoiding the additional complexity of such techniques, we consider $\hat{\mathbf{S}}^{\text{fix}}_M$ as a bivariate random sample of coordinates (n, m) for $n, m \in [1 : M]$ and our goal is to estimate the probability density function (pdf) producing this SSM.[3] The underlying assumption is that the pdf corresponds to the musical structure of the piece and that the bivariate random samples we observe are affected by variations in the realization of a specific performance. Analogous to [11], we employ a Parzen-Rosenblatt kernel density estimation method [13] that consist in convolving $\hat{\mathbf{S}}^{\text{fix}}_M$ with a two-dimensional Gaussian kernel of size ℓ. As a result, temporal variations in the performances are smoothed out. The choice of the value ℓ constitutes a trade-off between fingerprint characteristic (small value) and robustness to temporal variations (large value).

The resulting fingerprints (see Figure 1f) are an $M \times M$ representation[4] of the musical structure that features a high degree of robustness against properties of a specific performance. Figure 2 shows two further examples of fingerprints for the Mazurka Op. 56 No. 1.

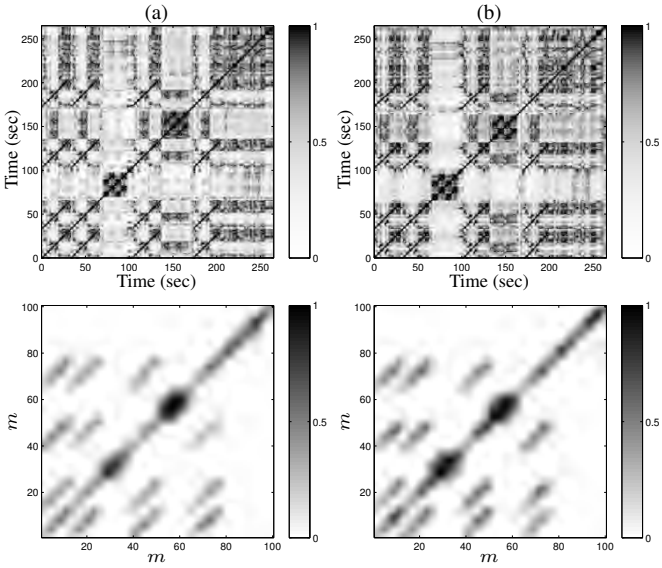

Figure 2: Original SSMs **(top)** and structure fingerprints **(bottom)** for two performances of Chopin's Mazurka Op. 56 No. 1. **(a)** Rubinstein (1966) and **(b)** Kushner (1989).

3. STRUCTURE-BASED RETRIEVAL

In this section, we show how the structure fingerprints (SF) can be used to facilitate structure-based music retrieval.

[2] In our experiments, using linear or cubic interpolation did not lead to any improvements.

[3] In the following, we use the term *pdf*, although for discrete random variables, the term *probability mass function* would be more appropriate.

[4] Note that this matrix is symmetric and only $M(M+1)/2$ entries are needed for representing the fingerprints.

Method	Dist.	Dataset	P	Sync.	MAP	T [sec]
Bello [1]	NCD	Bello	2919	No	0.767	>1000
SF	KL	ORG	2793	No	0.819	66.45
SF	ED	ORG	2793	No	0.816	0.58
SF	ED	MOD	2792	No	0.828	0.58
SF	ED	MOD	2792	Yes	0.958	0.58

Table 1: Overview of the results obtained for different methods and datasets. *Dist.* denotes the distance measure used, P the number of performances in the dataset, and T the run-time in seconds for computing $P \times P$ distances. [6] See Section 3.4 for a description of the dataset MOD and the column *Sync.* (indicating whether synchronized fingerprints are used).

We first describe the collection of Chopin Mazurkas (Section 3.1) and the retrieval scenario (Section 3.2). Then, we continue with a quantitative evaluation (Section 3.3) and give a musically informed discussion (Section 3.4).

3.1 Mazurka Collection

In our experiments, we use an audio collection comprising many recorded performances for each of the 49 Mazurka by Frédéric Chopin. Since different performances of a Mazurka typically share the same structure, this collection is a good choice for evaluating structural similarities. The dataset was assembled by the Mazurka Project [5] and has also been used by Bello in [1]. Note, however, that there are differences between our dataset (denoted as ORG in the following) and the one used in [1] (denoted as Bello). Actually, the datasets constitute a snapshot at different stages in the assembly process of the Mazurka Project which also results in a different number of performances (2793 for ORG and 2919 for Bello, see Table 1).

3.2 Retrieval Scenario

Using this dataset, we evaluate our structure fingerprints (SF) in a document-level retrieval scenario as in [1]. Given one performance of a Mazurka as query, the goal is to retrieve all other performances of the same Mazurka from the given dataset. To this end, we first compute the fingerprints for all P performances of the dataset. Using a suitable distance measure, we then derive the $P \times P$ matrix of pairwise distances between all performances, see Figure 6a. As the structure fingerprints are represented as densities, a natural choice of distance measure is the Kullback-Leibler divergence (KL). Additionally, in our experiments, we also use a simple Euclidean distance (ED). Finally, we rank the result with respect to ascending distances and express the retrieval accuracy by means of the *mean average precision* (MAP) measure as in [1, 12].

3.3 Quantitative Evaluation

First, we give a quantitative discussion of the results. Table 1 shows overall MAP values for the different methods and datasets. In [1], Bello reported MAP $= 0.767$ using his approach based on the NCD. Using the parameters $L = 10, \kappa = 20, M = 50, \ell = 5$ and the KL divergence, our approach leads to comparable, if not even slightly better results (MAP $= 0.819$). Note, however, that the results are not directly comparable due to the differences in the datasets. The results are insofar surprising, as our approach is not only conceptually much simpler, but also more explicit and, as it turns out, much more efficient. The last column of Table 1 indicates the run-time in seconds for computing the matrix of $P \times P$ pairwise distances. [6] Without knowing exact numbers for the NCD, our approach using KL seems to be at least one order of magnitude faster than [1]. Actually, when using the Euclidean distance (ED) instead of KL, the run-time of our approach can be improved significantly by two orders of magnitude (resulting in a run-time of just 0.58 seconds for computing all $P \times P$ distances), without any degradation of retrieval accuracy (MAP $= 0.816$).

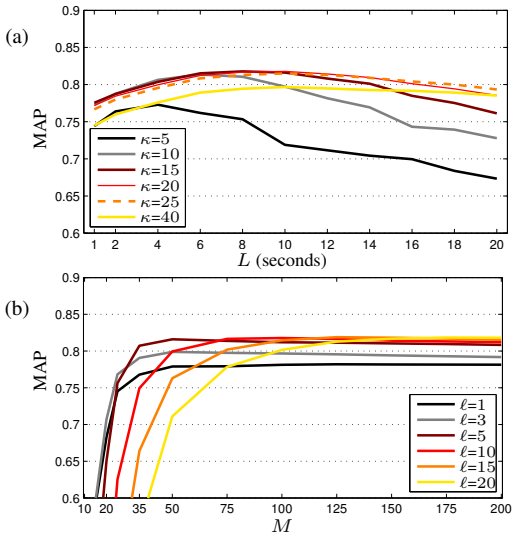

Figure 3: Parameter evaluation using MOD and Euclidean distances (ED). MAP values for different values of **(a)** the smoothing parameter L and threshold κ ($M = 100$, $\ell = 10$) and **(b)** of the fingerprint size M and kernel size ℓ ($L = 12$ sec, $\kappa = 20$).

We now continue with an evaluation of different parameter settings using ED (using KL lead to very similar findings). Figure 3a shows MAP values obtained on ORG as a function of the temporal smoothing parameter L (in seconds) and the relative threshold κ, see Section 2.2. Appropriate values for L constitute a trade-off between enhancement capability and level of detail. For the Mazurkas, a smoothing of 6-12 seconds seems to be reasonable, the actual choice of the parameter, however, is not crucial. For example, fixing $\kappa = 15$, one obtains MAP $= 0.815$ for $L = 6$ and MAP $= 0.816$ for $L = 10$. The threshold value κ constitutes a trade-off between retaining relevant structural information and denoising the SSMs. For the Mazurkas, 10%-25% seems to be a good compromise for capturing the repetitive structure. Again, the exact value is not crucial. For example, fixing $L = 6$, one obtains MAP $= 0.809$ for $\kappa = 25$ and MAP $= 0.814$ for $\kappa = 10$.

Figure 3b shows MAP values as a function of the fingerprint size M for different settings of the kernel density parameter ℓ. Interestingly, the size of the structure finger-

[5] mazurka.org.uk

[6] Using a vectorized MATLAB implementation of ED, a C/C++ implementation of KL, and an Intel Xeon E3-1225 CPU. Run-times for the NCD are estimated from the indicators given in [1] and own experiments.

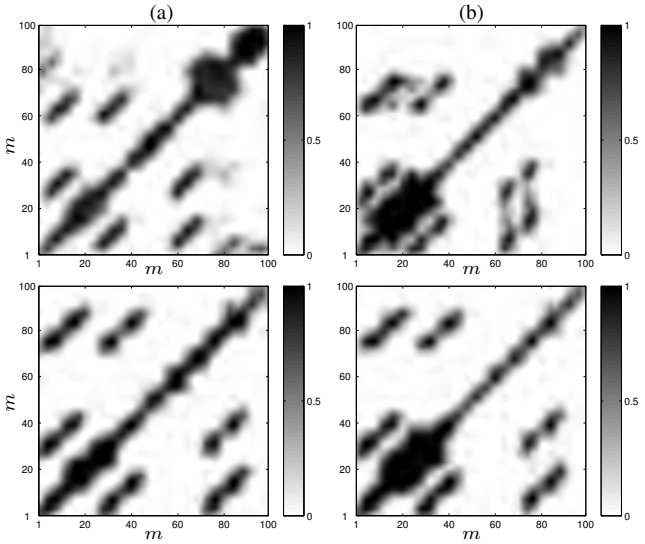

Figure 4: Structure fingerprints **(top)** and synchronized structure fingerprints **(bottom)** for performances of Chopin's Mazurka Op. 24 No. 2. **(a)** Merzhanov (2004) with applause at start and end of recording. **(b)** Smith (1975) with silence at the end.

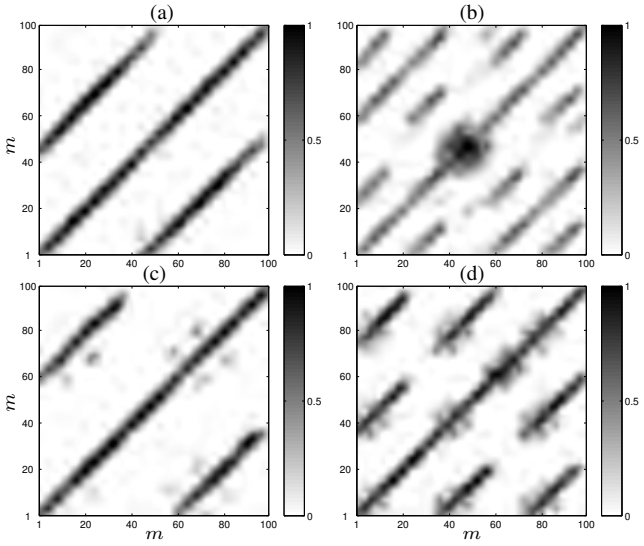

Figure 5: Structure fingerprints for four performances with differing structure of Chopin's Mazurka Op. 68 No. 4 ($L = 12$, $\kappa = 20$, $\ell = 10$, $M = 100$). **(a)** Niedzielski (1931). **(b)** Katin (1996). **(c)** Rubinstein (1952). **(d)** Rubinstein (1966).

prints can be reduced to $M = 50$ or even $M = 35$, while still retaining a high retrieval accuracy. The ratio of M and ℓ, however, is of crucial importance as it constitutes a trade-off between fingerprint characteristic and robustness against temporal variations in the performances. The settings $M = 50, \ell = 5$, and $M = 100, \ell = 10$, and $M = 200, \ell = 20$ yield almost identical retrieval results ($\mathrm{MAP} = 0.816$, $\mathrm{MAP} = 0.818$, and $\mathrm{MAP} = 0.819$, respectively). Decreasing the size of the fingerprints, however, has the advantage of reducing the computational load.

Aside from the robustness to actual parameter settings, our approach turned out to be rather robust to implementation details. For example, very similar results were obtained by using, e. g., Cosine, Hellinger, and Battacharyya distances between SFs. Even an alternative implementation using different chroma features as well as delay coordinates and recurrence plots (instead of the enhanced SSMs) similar to [1, 11, 12], lead to almost identical results. This also indicates that our concept is generalizable.

3.4 Musically Informed Discussion

Our fingerprint-based approach allows for detecting musically interesting phenomena and inconsistencies in the Mazurka collection. A careful investigation of the retrieval results revealed three phenomena. Firstly, we discovered that there are 67 recordings in the dataset that are incorrectly assigned to one of the Mazurkas, although they actually are performances of another Mazurka. [7] Another recording of the collection did not correspond to any of the Mazurkas. [8] We corrected these errors and denote the modified dataset `MOD`. Repeating the retrieval experiment using the 2792 performances of `MOD`, the MAP value increases to 0.828, see Table 1 (fourth row).

Secondly, it turned out that many incorrectly retrieved performances exhibit a long passages of applause, silence, or spoken moderation at the beginning and/or end. Actually, such passages can be regarded as additional structural elements. As a result, the structure of these performances does not match to the structure of the other performances of the same Mazurka, see Figure 4. To quantify this phenomenon, we use music synchronization techniques [3, 8] for identifying musically corresponding time positions in all versions of a Mazurka and use this information to warp the fingerprints to a common time line. For additional segments appearing in one performance, there are no corresponding time positions in the other performances. As a result, such segments are basically not reflected in the resulting *synchronized fingerprints*, see Figure 4. [9] Using synchronized fingerprints to exclude the additional segments, we repeat our experiment using `MOD` and obtain $\mathrm{MAP} = 0.958$, see Table 1 (last row).

The third phenomenon detected during our experiments are structural differences in the recordings. For instance, some pianists do not strictly stick to the score when performing a piece but omit (or sometimes even introduce) repetitions. Obviously, these structural differences lead to high distances as shown in Figure 6b for the Mazurka Op. 56 No. 1, where eight of the 42 performances exhibit a different structure. [10] The prime example for this effect is Mazurka Op. 68 No. 4, where the last bar in the score contains the marking *D. C. dal segno senza fine*. However, there is no *fine* marked in the score that would tell the pianist where to end. As a result, a performer may repeat the piece as often as he or she wants. This leads to many versions of the piece that differ significantly in structure as also revealed by the respective pairwise distances shown in Figure 6e. Figure 5 shows the fingerprints of four such

[7] A majority (51 of the 67 recordings) affects Op. 41 consisting of four Mazurkas (No. 1 to No. 4), where a permutation of the assigned numbers occurs.

[8] Labeled as a Rosenthal (1935) performance of Op. 50 No. 2.

[9] This strategy has a similar effect as using a distance measure based on dynamic programming, as proposed in [6, 7].

[10] Actually, all eight musicians omit a repetition of the A-part, leading to the form $A_1BA_2CA_3D$ instead of $A_1A_2BA_3CA_4D$.

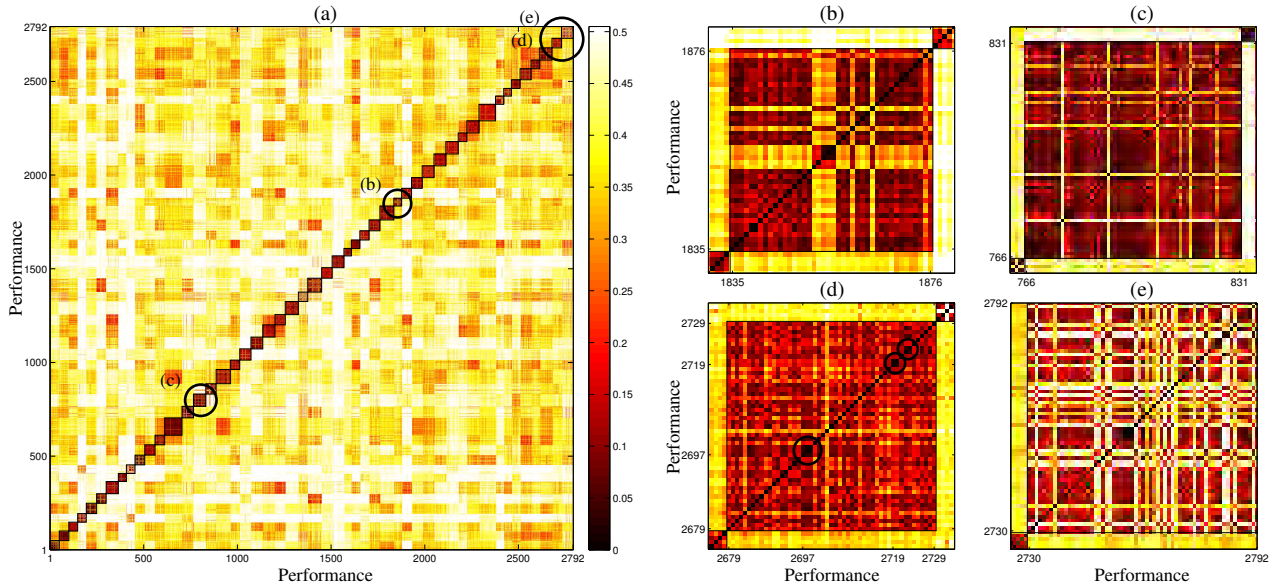

Figure 6: **(a)** Matrix of pairwise Euclidean distances for the 2792 performances of MOD. **(b)** Detail of the 42 performances of Op. 56 No. 1, see also Figure 2. **(c)** Detail of the 66 performances of Op. 24 No. 2, see also Figure 4. **(d)** Detail of the 51 performances of Op. 68 No. 3. **(e)** Detail of the 63 performances of Op. 68 No. 4, see also Figure 5.

versions, which, obviously, cannot be retrieved by a purely structure-based retrieval approach.

On the other hand, during our experiments we discovered performances that exhibit a surprisingly low distance, see, e. g., the squares of low distance on the main diagonal in Figure 6d. The low distance between the performances 2697-2699 is actually known as the "Hatto effect": recordings released under the name of the pianist Joyce Hatto in 1993 (2697) and 2006 (2698) that are actually time-scaled copies of a 1988 recordings of Eugen Indjic (2699). Similarly, some performances appear repeatedly in the dataset as they were released multiple times. Examples for this effect are performances 2719 and 2720 (Rubinstein) as well as 2722 and 2723 (Smidowicz).

4. CONCLUSION

The concept of structure fingerprints presented in this paper allows for retrieving music recordings solely based on structural information. Using a combination of suitable enhancement strategies, our approach is robust as well as efficient. Furthermore, as our experiments reveal, the results obtained by our approach are at least comparable to state-of-the-art approaches without relying on complex distance measures. As further advantage of our approach, just using Euclidean distances between fingerprints opens the possibility of exploiting efficient index-based methods such as locality-sensitive hashing to scale the approach to even larger datasets. We showed that our methods are suited for systematically analyzing structural properties of entire music collections, thus deepening the musical understanding of the data. Obviously, the limits of structure-based retrieval are reached when the assumption of global structural correspondence between performances is violated.

Acknowledgments: The work by P. Grosche und M. Müller has been supported by the Cluster of Excellence on Multimodal Computing and Interaction at Saarland University and the German Research Foundation (DFG MU 2686/5-1). J. Serrà and J. L. Arcos acknowledge 2009-SGR-1434 from Generalitat de Catalunya, TIN2009-13692-C03-01 from the Spanish Government, and EU Feder funds. J. Serrà also acknowledges JAEDOC069/2010 from Consejo Superior de Investigaciones Científicas.

5. REFERENCES

[1] J. P. Bello. Measuring structural similarity in music. *IEEE Trans. on Audio, Speech and Language Processing*, 19(7):2013–2025, 2011.

[2] M. A. Casey, R. Veltkap, M. Goto, M. Leman, C. Rhodes, and M. Slaney. Content-based music information retrieval: Current directions and future challenges. *Proc. of the IEEE*, 96(4):668–696, 2008.

[3] S. Ewert, M. Müller, and P. Grosche. High resolution audio synchronization using chroma onset features. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1869–1872, Taipei, Taiwan, 2009.

[4] J. Foote. ARTHUR: retrieving orchestral music by long-term structure. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Plymouth, USA, 2000.

[5] P. Grosche and M. Müller. Toward characteristic audio shingles for efficient cross-version music retrieval. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 473–476, Kyoto, Japan, 2012.

[6] T. Izumitani and K. Kashino. A robust musical audio search method based on diagonal dynamic programming matching of self-similarity matrices. In *Proc. of the International Conference on Music Information Retrieval (ISMIR)*, pages 609–613, Philadelphia, USA, 2008.

[7] B. Martin, M. Robine, and P. Hanna. Musical structure retrieval by aligning self-similarity matrices. In *Proc. of the Int. Conf. on Music Information Retrieval (ISMIR)*, pages 483–488, Kobe, Japan, 2009.

[8] M. Müller. *Information Retrieval for Music and Motion*. Springer Verlag, 2007.

[9] M. Müller and F. Kurth. Enhancing similarity matrices for music audio analysis. In *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 437–440, Toulouse, France, 2006.

[10] J. Paulus, M. Müller, and A. P. Klapuri. Audio-based music structure analysis. In *Proc. of the Int. Conf. on Music Information Retrieval (ISMIR)*, pages 625–636, Utrecht, The Netherlands, 2010.

[11] J. Serrà, M. Müller, P. Grosche, and J. L. Arcos. Unsupervised detection of music boundaries by time series structure features. In *Proc. of the AAAI Int. Conf. on Artificial Intelligence*, 2012. In Press.

[12] J. Serrà, X. Serra, and R. G. Andrzejak. Cross recurrence quantification for cover song identification. *New Journal of Physics*, 11(9):093017, 2009.

[13] J. S. Simonoff. *Smoothing Methods in Statistics*. Springer, 1996.