# RETRIEVING AUDIO RECORDINGS USING MUSICAL THEMES

*Stefan Balke, Vlora Arifi-Müller, Lukas Lamprecht, Meinard Müller*

International Audio Laboratories Erlangen, Friedrich-Alexander-Universität (FAU), Germany

{stefan.balke, meinard.mueller}@audiolabs-erlangen.de

## ABSTRACT

In 1948, Barlow and Morgenstern released a collection of about 10,000 themes of well-known instrumental pieces from the corpus of Western Classical music [1]. These monophonic themes (usually four bars long) are often the most memorable parts of a piece of music. In this paper, we report on a systematic study considering a cross-modal retrieval scenario. Using a musical theme as a query, the objective is to identify all related music recordings from a given audio collection. By adapting well-known retrieval techniques, our main goal is to get a better understanding of the various challenges including tempo deviations, musical tunings, key transpositions, and differences in the degree of polyphony between the symbolic query and the audio recordings to be retrieved. In particular, we present an oracle fusion approach that indicates upper performance limits achievable by a combination of current retrieval techniques.

***Index Terms***— Music Information Retrieval, Query-by-Example

## 1. INTRODUCTION

There has been a rapid growth of digitally available music data including audio recordings, digitized images of scanned sheet music, album covers, and an increasing number of video clips. The huge amount of readily available music requires retrieval strategies that allow users to explore large music collections in a convenient and enjoyable way. In the last decades, many systems for content-based audio retrieval scenarios that follow the query-by-example paradigm have been suggested. Given a fragment of a symbolic or acoustic music representation used as a query, the task is to automatically retrieve documents from a music database containing parts or aspects that are similar to the query [2–5]. One such retrieval scenario is known as *query-by-humming* [6, 7], where the user specifies a query by singing or humming a part of a melody. The objective is then to identify all audio recordings (or other music representations) that contain a melody similar to the specified query. Similarly, the user may specify a query by playing a characteristic phrase of a piece of music on an instrument [8, 9]. In a related retrieval scenario, the task is to identify an audio recording by means of a short symbolic query, e.g., taken from a musical score [10–12]. In the context of digital music libraries, content-based retrieval techniques are used to identify pieces in large archives which have not yet been systematically annotated [13, 14].

The retrieval scenario considered in this paper is inspired by the book "A Dictionary of Musical Themes" by Barlow and Morgenstern [1], which contains roughly 10,000 musical themes of instrumental Western classical music. Published in the year 1948, this dictionary is an early example of indexing music by its prominent themes. It was designed as a reference book for trained musicians and professional performers to identify musical pieces by a short query fragment. Most of the 10,000 themes listed in the book [1] are also available as machine-readable versions (MIDI) on the internet [15].

In this paper, we consider a cross-modal retrieval scenario, where the queries are symbolic encodings of musical themes and the database documents are audio recordings of musical performances. Then, given a musical theme used as a query, the task is to identify the audio recording of the musical work containing the theme. The retrieved documents may be displayed by means of a ranked list. This retrieval scenario offers several challenges.

- **Cross-modality.** On the one hand, we deal with symbolic sheet music (or MIDI), and with acoustic audio recordings on the other.
- **Tuning.** The tuning of the instruments, ensembles, and orchestras may differ from the standard tuning.
- **Transposition.** The key of a recorded performance may differ from the original key notated in the sheet music (e.g., transposed versions adapted to instruments or voices).
- **Tempo differences.** Musicians do not play mechanically, but speed up at some passages and slow down at others in order to shape a piece of music. This leads to global and local tempo deviations between the query fragments and the performed database recordings.
- **Polyphony.** The symbolic themes are monophonic. However, in the database recording they may appear in a polyphonic context, where the themes are often superimposed with other voices, countermelodies, harmonies, and rhythms.

Additionally, there can be variations in instrumentation, timbre, or dynamics. Finally, the audio quality of the recorded performances may be quite low, especially for old and noisy recordings.

The main motivation of this paper is to demonstrate the performance of standard music retrieval techniques that were originally designed for audio matching and version identification [16, Chapter 7]. By successively adjusting the retrieval pipeline, we perform an error analysis, gain a deeper understanding of the data to be matched, and indicate potential and limitations of current retrieval strategies. We think that this kind of error analysis using a baseline retrieval system is essential before approaching the retrieval problem by introducing more sophisticated and computationally expensive audio processing techniques, such as [9]. The remainder of the paper is structured as follows. In Section 2, we summarize the matching techniques and formalize the retrieval task. Then, in Section 3, we conduct extensive experiments and discuss our results. Further related work is discussed in the respective sections.

The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and the Fraunhofer-Institut für Integrierte Schaltungen IIS. This work has been supported by the German Research Foundation (DFG MU 2682/5-1).

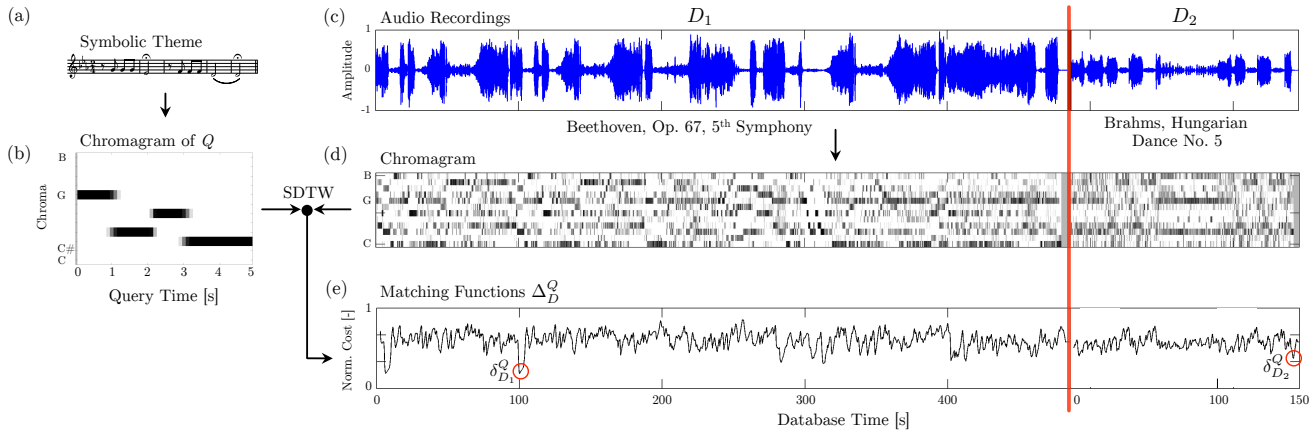


**Fig. 1**. Illustration of the matching procedure. **(a)** Sheet music representations of a musical theme. **(b)** Chromagram of the query. **(c)** Music collection as a concatenated waveform. **(d)** Chroma representation of the recordings in the music collection. **(e)** Matching function $\Delta$.

## 2. MATCHING PROCEDURE

In this section, we summarize the retrieval procedure used here, following [16]. Similar procedures for synchronizing polyphonic sheet music and audio recordings were described in the literature [10, 12].

### 2.1. Chroma Features

Chroma features have been successfully used in solving different music-related search and analysis tasks [16, 17]. These features strongly correlate with tonal (harmonic, melodic) components for music whose pitches can be meaningfully categorized (often into 12 chromatic pitch classes) and whose tuning approximates to the equal-tempered scale [18]. In particular, chroma features are suited to serve as a mid-level feature representation for comparing and relating acoustical and symbolic music, see Figure 1b and Figure 1d.

In our experiments (Section 3), we use the *Chroma Toolbox* [19] which uses a filterbank to decompose the audio signal in the aforementioned pitch classes. In particluar, we use a chroma feature variant called CENS features. Starting with a feature rate of 10 Hz, we apply a temporal smoothing over nine frames and a downsampling by a factor of two. This results in chroma features at a rate of 5 Hz, as used in our experiments (Section 3).

### 2.2. Matching Technique

To compare a symbolic query to an audio recording contained in a music collection, we convert the query and recording into chroma sequences, say $X := (x_1, x_2, \ldots, x_N)$ and $Y := (y_1, y_2, \ldots, y_M)$. Typically, the length $M \in \mathbb{N}$ of $Y$ is much larger than the length $N \in \mathbb{N}$ of the query $X$. Then, we use a standard technique known as *Subsequence Dynamic Time Warping* (SDTW) to compare $X$ with subsequences of $Y$, see [20, Chapter 4]. In particular, we use the cosine distance (for comparing normalized chroma feature vectors) and the step size condition $\Sigma_1 := \{(1,0),(0,1),(1,1)\}$ in the SDTW. Furthermore, for the three possible step sizes, one may use additional weights $w_v$, $w_h$, and $w_d$, respectively. In the standard procedure, the weights are set to $w_v = w_h = w_d = 1$. In our later experiments, we use the weights to further penalize certain steps. As the results of SDTW, one obtains a matching function $\Delta : [1 : M] \to \mathbb{R}$. Local minima of $\Delta$ point to locations with a good match between the query $X$ and a subsequence of $Y$, as indicated by the red circle in Figure 1e. For the details of this procedure and its parameters, we refer to [20, Chapter 4].

### 2.3. Retrieval Task

In the following, we formalize our retrieval task. Let $\mathcal{Q}$ be a collection of musical themes, where each element $Q \in \mathcal{Q}$ is regarded as a *query*. Furthermore, let $\mathcal{D}$ be a set of audio recordings, which we regard as a database collection consisting of *documents* $D \in \mathcal{D}$. Given a query $Q \in \mathcal{Q}$, the retrieval task is to identify the semantically corresponding documents $D \in \mathcal{D}$. In this setting, we are only interested in the associated audio recording of a given theme and not in its exact position within the recording. Therefore, we compute a matching function $\Delta_D^Q$ for $Q$ and each of the documents $D \in \mathcal{D}$. Then, we define $\delta_D^Q = \min_m \Delta_D^Q(m)$ to be the distance between $Q$ and $D$. Finally, we sort the database documents $D \in \mathcal{D}$ in ascending order according to the values $\delta_D^Q$. The position of a document $D$ in this ordered list is called the *rank* of $D$.

Figure 1 illustrates the matching procedure by using Beethoven's "Fate-Motif" as query. First, the given sheet music is transformed into a sequence of chroma features (see Figure 1a-b). In this example, our database consists of two audio recordings (see Figure 1c), which are also converted into chroma-based feature sequences (see Figure 1d). The matching functions $\Delta_D^Q$ are shown in Figure 1e. Red circles indicate the positions of the minima $\delta_D^Q$ for each document $D$. In this example, the matching function yields two distinct minima in the first document (Beethoven) at the beginning and after roughly 100 s. This is due to the fact that the motif, which is used as query, occurs several times in this work. In our document level scenario, both minima are considered to be correct matches as we are only interested in the entire recording and not in the exact position of the queried theme.

## 3. EXPERIMENTS

We now report on our experiments using queries from the book by Barlow and Morgenstern, where we successively adapt the described matching procedure. Our main motivation is to gain a better understanding of the challenges regarding musical tuning, key transpositions, tempo deviation, and the degree of polyphony.

### 3.1. Test Datasets

The symbolic queries as given in the book by Barlow and Morgenstern [1] are available on the internet as MIDI files [15] in the "Electronic Dictionary of Musical Themes" (in the following referred to as

| Queries | #Themes | Database | #Recordings | Duration |
|---|---|---|---|---|
| $\mathcal{Q}_1$ | 177 | $\mathcal{D}_1$ | 100 | ∼11 h |
| $\mathcal{Q}_2$ | 2046 | $\mathcal{D}_2$ | 1113 | ∼120 h |

**Table 1**. Overview of the datasets used for our experiments.

EDM). We denote the 9803 themes from EDM by $\mathcal{Q}$. Furthermore, let $\mathcal{D}$ be a collection of audio recordings $D \in \mathcal{D}$.

We created two query test datasets, as shown by Table 1. The first dataset $\mathcal{Q}_1$ consists of 177 queries and serves as a development testset. The second test dataset $\mathcal{Q}_2$ contains 2046 queries and is used to investigate the scalability of the matching technique. In both test datasets, the durations of the queries ranges roughly between 1 s and 19 s with a mean of 7.5 s.

Additionally, we design two collections $\mathcal{D}_1$ and $\mathcal{D}_2$, which contain exactly one audio recording representing a true match of the queries contained in $\mathcal{Q}_1$ and $\mathcal{Q}_2$, respectively. Note that the number of queries is higher than the number of recordings because for a given musical piece, several themes may be listed in the book by Barlow and Morgenstern; e.g., there are six musical themes listed for the first movement of Beethoven's 5th Symphony.

### 3.2. Evaluation Measures

In our evaluations, we compare a query $Q \in \mathcal{Q}$ with each of the documents $D \in \mathcal{D}$. This results in a ranked list of the documents $D \in \mathcal{D}$, where (due to the design of our test datasets $\mathcal{D}_1$ and $\mathcal{D}_2$) one of these documents is considered relevant. Inspired by a search-engine-like retrieval scenario, where a user typically looks at the top match and then may also check the first five, ten or twenty matches, we evaluate the top $K$ matches for $K \in \{1, 5, 10, 20\}$. For a given $K$, the query is considered to be correct if its retrieved rank is at most $K$. Considering all queries at question, we then compute the proportion of correct queries (w.r.t. $K$). This results in a number $\rho_K \in [0 : 100]$ (given in percent), which we refer to as Top-K matching rate. Considering different values for $K$ gives us insights in the distribution of the ranks and the system's retrieval performance.

### 3.3. Experiments using $\mathcal{Q}_1$ and $D_1$

We start with a first series of experiments based on $\mathcal{Q}_1$ and $\mathcal{D}_1$, where we systematically adapt various parameter settings while reducing the retrieval task's complexity by exploiting additional knowledge. We then aggregate the obtained results by means of an oracle fusion. This result indicates the upper limit for the performance that is achievable when using the suggested matching pipeline. Table 2 gives an overview of the results, which we now discuss in detail by exemplarily considering the results for $\rho_1$ and $\rho_{10}$.

**Baseline.** As a preliminery experiment, we use $\Sigma_1$ for the step size condition and $w_v = w_h = w_d = 1$ as weights. This yields Top-K matching rates of $\rho_1 = 38.4\,\%$ and $\rho_{10} = 62.7\,\%$. To increase the system's robustness, we restrict the SDTW procedure by using a different step size condition $\Sigma$. In general, using the set $\Sigma_1$ may lead to alignment paths that are highly deteriorated. In the extreme case, the query $X$ may be assigned to a single element of $Y$. Therefore, it may be beneficial to replace $\Sigma_1$ with the set $\Sigma_2 = \{(2,1),(1,2),(1,1)\}$, which yields a compromise between a strict diagonal matching (without any warping, $\Sigma_0 = \{(1,1)\}$) and the DTW-based matching with full flexibility (using $\Sigma_1$). Furthermore, to avoid the query $X$ being matched against a very short subsequence of $Y$, we set the weights to $w_v = 2$, $w_h = 1$, and $w_d = 1$. Similar settings have been used, e. g., in [21]. With these settings, we slightly improve the Top-K matching rates to $\rho_1 = 45.2\,\%$ and $\rho_{10} = 70.1\,\%$ (see also "Baseline" in Table 2). In the following, we continue using $\Sigma_2$ and the weights $w_v = 2$, $w_h = 1$, and $w_d = 1$.

| Top-K | 1 | 5 | 10 | 20 |
|---|---|---|---|---|
| Baseline | 45.2 | 62.1 | 70.1 | 76.8 |
| Tu | 46.9 | 64.4 | 72.9 | 81.9 |
| Tr | 52.0 | 68.9 | 79.1 | 87.6 |
| Tu+Tr | 53.7 | 72.3 | 83.1 | 91.0 |
| Tu+Tr+Ql | 68.4 | 79.1 | 88.1 | 93.2 |
| Tu+Tr+Ql+Df | 37.3 | 57.6 | 67.8 | 74.6 |
| Oracle Fusion | 72.3 | 84.7 | 92.1 | 97.7 |

**Table 2**. Top-K matching rate for music collection $\mathcal{D}_1$ with corresponding musical themes $\mathcal{Q}_1$ used as queries. The following settings are considered: Tu = Tuning estimation, Tr = Annotated transposition, Ql = Annotated query length, Df = Dominant feature band.

**Tuning (Tu) and Transposition (Tr).** Deviations from the standard tuning in the actual music recording can lead to misinterpretations of the measured pitch. Estimating the tuning used in the music recording beforehand can reduce these artifacts [17]. Instead of using a dedicated tuning estimator, we simply test three different tunings by detuning the filterbank by $\pm 1/3$ semitones used to compute the chroma features (see Section 2.1). We then pick the tuning which yields the smallest minimum $\delta_D^Q$. For a detailed descripton of a similar procedure, we refer to [17, 22]. This further improves the matching rates to $\rho_1 = 46.9\,\%$ and $\rho_{10} = 72.9\,\%$. As the musical key of the audio recording may differ from the key specified in the MIDI, we manually annotated the required transposition. Using this information in the matching procedure (by applying suitable chroma shifts [23]), the results improve to $\rho_1 = 52.0\,\%$ and $\rho_{10} = 79.1\,\%$. Combining both, the tuning estimation and the correct transposition, we get Top-K matching rates of $\rho_1 = 53.7\,\%$ and $\rho_{10} = 83.1\,\%$.

**Query Length (Ql).** We observed that the tempo events in some of our MIDI queries are set to an extreme parameter, which results in a query duration that strongly deviates from the corresponding passage in the audio recording. When the tempo information deviates too much from the audio recording, SDTW based on $\Sigma_2$ is unable to warp the query to the corresponding audio section. Furthermore, the features may lose important characteristics. For instance, the beginning theme of Beethoven's Pathétique has a MIDI duration of 3.5 s, whereas the corresponding section in the audio recording has a duration of 21 s. To even out tempo differences, we manually annotated the durations of the audio sections corresponding to queries and used this information to adapt the duration of the query before calculating the chroma features. This further increases the matching rate to $\rho_1 = 68.4\,\%$ and $\rho_{10} = 88.1\,\%$.

**Dominant Feature Band (Df).** In the next experiment, we want to compensate for the different degrees of polyphony. Looking at the chromagram of the monophonic musical theme in Figure 1b reveals that only one chroma band is active at a time. For database documents as shown in Figure 1d, however, the energy is spread across several chroma bands due to the instruments' partials and accompaniments. A first method to reduce the polyphony on the audio side is to only take the dominant chroma band (the band with the largest value) for each time frame. This can be thought of as "monofying" the database document in the mid-level feature representation. Using this monofied chroma representation results in a matching rate

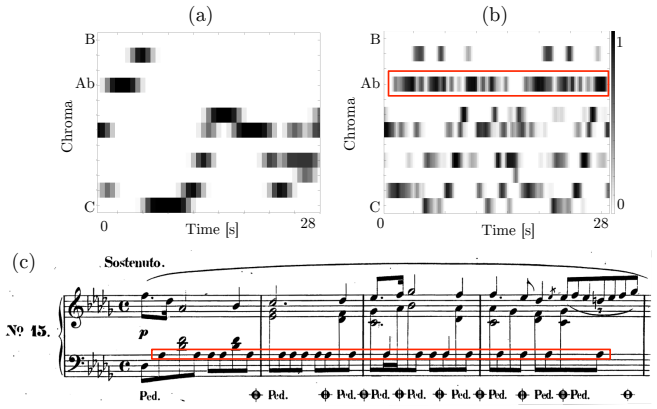


**Fig. 2**. Example of Chopin's Prélude Op. 28, No. 15 ("Raindrop"). **(a)** Chromagram of monophonic query. **(b)** Chromagram of the corresponding section in the audio recording. **(c)** Sheet music representation of the corresponding measures.

of $\rho_1 = 37.3\,\%$ and $\rho_{10} = 67.8\,\%$. Even though this procedure works for some cases, for others it may pick the "wrong" chroma band, thus deteriorating the overall retrieval result. Further experiments showed that more refined methods (by extracting the predominant melody as described in [24]), may lead to slightly better results. However, Figure 2a shows a typical example where the advanced methods still fail, since the salient energy is located in the A♭-band (see Figure 2b), which is the accompaniment played with the left hand (see Figure 2c) and not the part we would perceive as being the main melody.

**Oracle Fusion.** In this experiment we assume having an oracle which can tell us, for each query, which setting performs best (in the sense that the relevant document is ranked better). The results obtained from oracle fusion yield a kind of upper limit which can be reached by using the suggested matching pipeline. Performing the oracle fusion for all queries leads to matching rates of $\rho_1 = 72.3\,\%$ and $\rho_{10} = 92.1\,\%$ (see Table 2). Oracle fusion shows that our matching pipeline may yield good retrieval results. However, a good prior estimate of transposition and tempo is important. Also, as we see in our next experiment, the results do not scale well when considering much larger datasets.

### 3.4. Experiments using $\mathcal{Q}_2$ and $\mathcal{D}_2$

We now expand the experiments using the larger datasets $\mathcal{Q}_2$ (consisting of 2046 musical themes) and $\mathcal{D}_2$ (consisting of 1113 audio recordings). In this case, we do not have any knowledge of transposition and tempo information. One strategy to cope with different transpositions is to simply try out all 12 possibilities by suitably shifting the queries' chromagrams [23]. This, however, also increases the chance of obtaining false positive matches. Analyzing the annotations from $\mathcal{D}_1$, it turns out that most of the transpositions lie within $[-2:2]$ semitones. Therefore, in subsequent experiments, we only use these five transpositions, instead of all twelve possible chroma shifts. As for the query length, the durations of the annotated sections in $\mathcal{D}_1$ are within 3 s and 30 s. To cover this range, the duration of each query (EDM MIDI) is set to 5 s, 10 s, and 15 s, respectively. The results of the Top-K matching rates are shown in Table 3. For example, when using a query length of 5 s, the the matching rates are $\rho_1 = 14.9\,\%$ and $\rho_{10} = 25.8\,\%$. Using different query lengths (10 s and 15 s) does not substantially improve the retrieval results. However, using an oracle fusion over the different query lengths, the retrieval results substantially improve, leading to matching rates of $\rho_1 = 25.0\,\%$ and $\rho_{10} = 39.0\,\%$. In other words, even when using alignment methods to compensate for local tempo differences, a good initial estimate for the query duration is an essential step to improve the matching results.

| **Top-K** | **1** | **5** | **10** | **20** | **50** | **100** | **200** | **500** |
|---|---|---|---|---|---|---|---|---|
| Tu+Tr+5 s | 14.9 | 21.8 | 25.8 | 29.2 | 35.5 | 43.0 | 54.1 | 76.1 |
| Tu+Tr+10 s | 18.3 | 25.1 | 28.3 | 32.6 | 38.7 | 46.1 | 56.1 | 76.2 |
| Tu+Tr+15 s | 13.6 | 19.5 | 22.7 | 26.1 | 31.6 | 38.9 | 49.7 | 72.4 |
| Oracle Fusion | 25.0 | 34.1 | 39.0 | 43.5 | 51.0 | 59.6 | 70.2 | 86.9 |

**Table 3**. Top-K matching rate for music collection $\mathcal{D}_2$ with corresponding musical themes $\mathcal{Q}_2$ used as queries. The following settings are considered: Tu = Tuning estimation, Tr = Annotated transposition, $\{5, 10, 15\}$ s = Fixed query length.

Concluding these experiments, one can say that the retrieval of audio recordings by means of short monophonic musical themes is a challenging problem due to the challenges listed in the introduction (Section 1). We have seen that a direct application of a standard chroma-based matching procedure yields reasonable results for roughly half of the queries. However, the compensation of tuning issues and tempo differences is of major importance. The used matching procedure is simple to implement and has the potential for applying indexing techniques to speed up computations [25].

Differences in the degree of polyphony remain one main problem when matching monophonic themes against music recordings. In this context, simply taking the dominant feature band, as in our experiment in Section 3.3, turned out to even worsen the matching quality. (This was also the reason why we did not used this strategy in our experiment of Section 3.4.) One promising approach, as suggested in [9], is to use NMF-based techniques to decompose the audio recording into monophonic-like components. These techniques, however, are computationally expensive and do not easily scale to recordings of long duration and large datasets. The development of scalable techniques to match monophonic and polyphonic music representations remain a research direction with many challenging problems.

## 4. CONCLUSION AND FUTURE WORK

In this paper, we have presented some baseline experiments for identifying audio recordings by means of musical themes. Due to musical and acoustic variations in the data as well as the typically short duration of the query, the matching task turned out to be quite challenging. Besides gaining some deeper insights into the challenges and underlying data, we still see potential of the considered retrieval techniques—in particular within a cross-modal search context. For example, in the case of the Barlow–Morgenstern scenario, the book contains textual specifications of the themes besides the visual score representations of the notes. Similarly, structured websites (e.g., Wikipedia websites) often contain information of various types including text, score, images, and audio. By exploiting multiple types of information sources, fusion strategies may help to better cope with uncertainty and inconsistency in heterogeneous data collections (see [26]). For example, in [27], such a fusion approach was presented for identifying musical themes (given in MIDI format) based on corrupted OMR and OCR input. The further investigation of such cross-modal fusion approaches, including audio, image, and text-based cues, constitutes a promising research direction.

## 5. REFERENCES


[1] Harold Barlow and Sam Morgenstern, *A Dictionary of Musical Themes*, Crown Publishers, Inc., revised edtion, 3. edition, 1975.

[2] Michael A. Casey, Remco Veltkap, Masataka Goto, Marc Leman, Christophe Rhodes, and Malcolm Slaney, "Content-based music information retrieval: Current directions and future challenges," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 668–696, 2008.

[3] Peter Grosche, Meinard Müller, and Joan Serrà, "Audio content-based music retrieval," in *Multimodal Music Processing*, Meinard Müller, Masataka Goto, and Markus Schedl, Eds., vol. 3 of *Dagstuhl Follow-Ups*, pp. 157–174. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2012.

[4] Colin Raffel and Daniel P. W. Ellis, "Large-scale content-based matching of MIDI and audio files," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Málaga, Spain, 2015, pp. 234–240.

[5] Rainer Typke, Frans Wiering, and Remco C Veltkamp, "A survey of music information retrieval systems.," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, London, UK, 2005, pp. 153–160.

[6] Matti Ryynänen and Anssi Klapuri, "Query by humming of MIDI and audio using locality sensitive hashing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, 2008, pp. 2249–2252.

[7] Justin Salamon, Joan Serrà, and Emilia Gómez, "Tonal representations for music retrieval: from version identification to query-by-humming," *International Journal of Multimedia Information Retrieval*, vol. 2, no. 1, pp. 45–58, 2013.

[8] Andreas Arzt, Sebastian Böck, and Gerhard Widmer, "Fast identification of piece and score position via symbolic fingerprinting," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2012, pp. 433–438.

[9] Taro Masuda, Kazuyoshi Yoshii, Masataka Goto, and Shigeo Morishima, "Spotting a query phrase from polyphonic music audio signals based on semi-supervised nonnegative matrix factorization," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2014, pp. 227–232.

[10] Christian Fremerey, Michael Clausen, Sebastian Ewert, and Meinard Müller, "Sheet music-audio identification," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Kobe, Japan, Oct. 2009, pp. 645–650.

[11] Jeremy Pickens, Juan Pablo Bello, Giuliano Monti, Tim Crawford, Matthew Dovey, Mark Sandler, and Don Byrd, "Polyphonic score retrieval using polyphonic audio," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2002.

[12] Iman S.H. Suyoto, Alexandra L. Uitdenbogerd, and Falk Scholer, "Searching musical audio using symbolic queries," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 16, no. 2, pp. 372–381, 2008.

[13] David Damm, Christian Fremerey, Verena Thomas, Michael Clausen, Frank Kurth, and Meinard Müller, "A digital library framework for heterogeneous music collections: from document acquisition to cross-modal interaction," *International Journal on Digital Libraries: Special Issue on Music Digital Libraries*, vol. 12, no. 2-3, pp. 53–71, 2012.

[14] Nicola Montecchio, Emanuele Di Buccio, and Nicola Orio, "An efficient identification methodology for improved access to music heritage collections," *Journal of Multimedia*, vol. 7, no. 2, pp. 145–158, 2012.

[15] Jacob T. Schwartz and Diana Schwartz, "The electronic dictionary of musical themes," Website `http://www.multimedialibrary.com/barlow/`, last accessed 01/12/2015, 2008.

[16] Meinard Müller, *Fundamentals of Music Processing*, Springer Verlag, 2015.

[17] Emilia Gómez, *Tonal Description of Music Audio Signals*, Ph.D. thesis, UPF Barcelona, 2006.

[18] Carol L. Krumhansl, *Cognitive foundations of musical pitch*, Oxford University Press, 1990.

[19] Meinard Müller and Sebastian Ewert, "Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Miami, Florida, USA, 2011, pp. 215–220.

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

[21] Meinard Müller and Sebastian Ewert, "Towards timbre-invariant audio features for harmony-based music," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 649–662, 2010.

[22] Meinard Müller, Peter Grosche, and Frans Wiering, "Robust segmentation and annotation of folk song recordings," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Kobe, Japan, Oct. 2009, pp. 735–740.

[23] Masataka Goto, "A chorus-section detecting method for musical audio signals," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, China, 2003, pp. 437–440.

[24] Justin Salamon and Emilia Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1759–1770, 2012.

[25] Peter Grosche and Meinard 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)*, Kyoto, Japan, 2012.

[26] Meinard Müller, Masataka Goto, and Markus Schedl, Eds., *Multimodal Music Processing*, vol. 3 of *Dagstuhl Follow-Ups*. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Germany, 2012.

[27] Stefan Balke, Sanu Pulimootil Achankunju, and Meinard Müller, "Matching musical themes based on noisy OCR and OMR input," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane, Australia, 2015.