MATCHING MUSICAL THEMES BASED ON NOISY OCR AND OMR INPUT

Stefan Balke, Sanu Pulimootil Achankunju, Meinard Müller

International Audio Laboratories Erlangen, Friedrich-Alexander-Universität (FAU), Germany
{stefan.balke, meinard.mueller}@audiolabs-erlangen.de

ABSTRACT

In the year 1948, Barlow and Morgenstern published the book "A Dictionary of Musical Themes", which contains 9803 important musical themes from the Western classical music literature. In this paper, we deal with the problem of automatically matching these themes to other digitally available sources. To this end, we introduce a processing pipeline that automatically extracts from the scanned pages of the printed book textual metadata using Optical Character Recognition (OCR) as well as symbolic note information using Optical Music Recognition (OMR). Due to the poor printing quality of the book, the OCR and OMR results are quite noisy containing numerous extraction errors. As one main contribution, we adjust alignment techniques for matching musical themes based on the OCR and OMR input. In particular, we show how the matching quality can be substantially improved by fusing the OCR- and OMR-based matching results. Finally, we report on our experiments within the challenging Barlow and Morgenstern scenario, which also indicates the potential of our techniques when considering other sources of musical themes such as digital music archives and the world wide web.

Index Terms— Music Information Retrieval, Optical Character Recognition, Optical Music Recognition, 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 [1, 2, 3, 4, 5]. In this paper, we focus on Western classical music, where a piece of music is typically specified by the composer, some work identifier such as a catalogue or opus number, and other types of metadata. For example, the musical work number Op. 67 by Ludwig van Beethoven specifies his Symphony No. 5 in C minor, the symphony with the famous fate motive. Besides such textual descriptions, Western classical music is given in form of printed sheet music, which visually encodes the notes to be played by musicians. Thanks to massive digitization efforts like the International Music Score Library Project[1] (IMSLP), millions of digitized pages of sheet music are publicly available on the world wide web.

Handling music collections of this size, one requires analysis and retrieval techniques for the various kinds of representations and formats. One important step consists in extracting the textual metadata as well as the note information from the digitized images. To this end, techniques such as Optical Character Recognition (OCR) to extract text-based metadata and Optical Music Recognition (OMR) to extract symbolic representations from the digital scans of printed sheet music are needed [6, 7, 8, 9, 10]. Besides of inconsistencies in the metadata that describes a musical work, the OCR and OMR may contain a significant number of extraction errors. This particularly holds for books of poor printing quality and scans of low resolution.

In this paper, we deal with a challenging matching scenario by considering the book "A Dictionary of Musical Themes" by Barlow and Morgenstern [11]. This book yields an overview of the most important musical themes from the Western classical music literature, thus covering many of the pieces contained in IMSLP. The contributions of this paper are as follows. First, we describe a fully automated processing pipeline that matches the music themes from the book by Barlow and Morgenstern to other digitally available sources. This pipeline involves segmentation, OCR, OMR, and alignment techniques (see Section 2 and Fig. 1). Then, we report on extensive experiments that indicate the retrieval quality based on inconsistent and erroneous OCR and OMR input (see Section 3). In particular, we show how the quality can be significantly improved by fusing the OCR-based and OMR-based matching results. Finally, we discuss how our processing pipeline may be applied to automatically identify, retrieve, and annotate musical sources that are distributed in digital music archives and the world wide web.

2. PROCESSING PIPELINE

2.1. Text and Score Recognition

As starting point for our matching scenario, we use the book by Barlow and Morgenstern [11], which contains 9803 musical themes from the most important compositions of the Western classical music literature. The book includes orchestral music, chamber music, and works for solo instruments. Each theme is specified by a textual specification as well as a visual score representation of the notes. In particular, the respective composer, the underlying musical work, and the movement are listed. Within the book, the themes are systematically organized and suitably indexed.

An example for a scanned page of the book is shown in Fig. 1a. The excerpt shows text-based metadata as well as score information. The composer is written on the top of each page (e. g., "Beethoven"), whereas the title of each musical work (e. g., "Symphony No. 5 in C Minor") is specified in a text box aligned to the left. Furthermore, each theme is further specified by a movement and theme description (e. g., "1st Movement, 1st Theme, A") followed by a score representation of the theme. Finally, an additional identifier (e. g., "B948"), which is used for indexing purposes, is printed at the end of each theme.

As this example shows, the book is structured in a systematic fashion, even though the positions of the various text boxes may

The International Audio Laboratories Erlangen is a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and the Fraunhofer-Institut für Integrierte Schaltungen IIS.

[1] http://www.imslp.org/

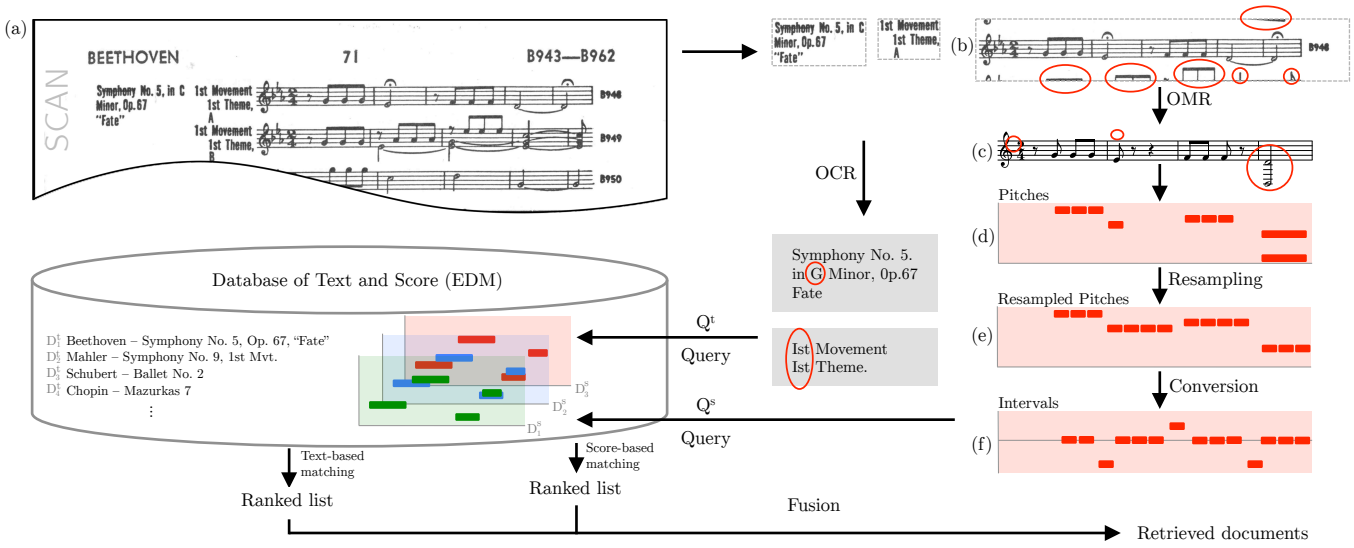

Fig. 1. Overview of the processing pipeline. Each page is segmented into text and sheet music parts. The cropped images are transformed into computer readable representations using OCR and OMR (typical extraction errors are highlighted by a red circle). The results are used to query against a database consisting of music documents. Using a fusion strategy based on text-based and score-based matching results, the retrieval system outputs a ranked list of documents.

slightly vary from theme to theme and page to page. Using heuristics on the layout of the book, we first automatically segment each page by determining for each of the themes the bounding boxes of the various text elements and the image containing the score information. In particular, we exploit the knowledge on the rough position of the elements as well as the characteristic horizontal staff lines of the score. This yields a segmentation result as indicated in Fig. 1b. Because of the regular structure of the pages, the bounding boxes computed by our algorithm are correct for more than 99 % of the themes. One problem is that the bounding boxes for the score representations may intersect with previous and subsequent bounding boxes, which often results in unwanted score fragments as highlighted in Fig. 1b.

The text boxes are further processed by feeding in the cropped images into an OCR engine. In our processing pipeline, we have used the freely available OCR engine Tesseract [12]. As indicated by Fig. 1, the recognition results are of good overall quality with occasional errors on the character level. In our example, the string "1st" has been recognized as "Ist" and "C Minor" was transcribed as "G Minor". Because of its prominent placement, the larger font size, and the capitalization, the extraction of the composers' names (e. g., "BEETHOVEN") works particularly well.

The score information is processed by feeding in the cropped images into an OMR engine. For this task, we use the freely available OMR software Audiveris [13]. As can be seen by our example, the score conversion is more problematic than in the case of text. On the one hand, many extraction errors occur on the note level. In our example, some of the note lengths were not detected correctly, the fermata is missing, and an additional note has been added in the last measure. Some of these errors come from score fragments due to the above mentioned intersection problem of the bounding boxes. On the other hand, there are recognition errors that have a global impact on the interpretation of the pitch parameters of the notes. In particular, the recognition of the key and time signatures as well as the kind of clef (e.g. G-clef, C-clef or F-clef) has turned out to be problematic. In the example of Fig. 1c, the OMR engine could not detect the three flats of the key signature, which affects the interpretation of the fourth note (the E flat becomes an E). Most of the errors are due to the poor printing quality of the book by Barlow and Morgenstern. Experiments with different scan resolutions and other OMR engines (e. g., PhotoScore, SharpEye or SmartScore) have not resolved these problems. As we will show in the next section, the influence of the extraction errors can be attenuated by designing suitable cost functions and matching procedures.

2.2. Matching Procedures

As a result of the previously described recognition process, we obtain a textual representation of the metadata (containing the composer, work identifier, and other metadata) and a symbolic score representation for each of the 9803 themes of the book by Barlow and Morgenstern (in the following referred to as BM). The goal is to use this information for identifying other digital sources that belong or relate to the musical themes. In our experiments, we consider a scenario that allows us to study various matching procedures and to systematically evaluate matching results. To this end, we consider the "Electronic Dictionary of Musical Themes" (in the following referred to as EDM), which is publicly available at [14]. The EDM collection contains Standard MIDI files for the musical themes, which are linked to textual metadata similar to the original book by Barlow and Morgenstern. While the EDM themes more or less agree with the BM themes, there are inconsistencies with regard to the number of themes, the metadata and the score representations. Using the printed BM book as a reference, we have manually linked the BM themes to corresponding EDM themes. These correspondences serve as ground truth in the subsequent experiments.

In the following, we formulate our setting as a retrieval task. We denote the set of BM themes by $\mathcal{Q}$, where each element $Q \in \mathcal{Q}$ is regarded as a *query*. Furthermore, let $\mathcal{D}$ be the set of EDM themes, 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 document $D \in \mathcal{D}$.

2.3. Text-based Matching

Let us consider a fixed query $Q \in \mathcal{Q}$. In a first matching procedure, we only consider the textual representation, denoted by Q^{t}, which was obtained from the OCR step. Similarly, let D^{t} denote the text information for a document $D \in \mathcal{D}$. Both Q^{t} as well as D^{t} are represented as character strings. To compare these strings, one can use standard string alignment techniques such as the edit distance [15]. In our scenario, the two strings to be compared both contain the name of the composer, some work descriptor as well as a movement and theme identifier. However, the strings may also differ substantially due to additional information, segmentation errors, and OCR errors. Therefore, to compare strings, we use the longest common subsequence (LCS), which is a variant of the edit distance that is more robust to noise and outliers. For a description of this standard similarity measure, we refer to [15]. We convert the LCS-based similarity value into a normalized cost value by defining

$$c^{\mathrm{t}}(Q, D) := 1 - \frac{\mathrm{LCS}(Q^{\mathrm{t}}, D^{\mathrm{t}})}{|Q^{\mathrm{t}}|} \in [0, 1], \tag{1}$$

where $|Q^{\mathrm{t}}|$ denotes the length of the string Q^{t}. The performance of this matching procedure is discussed in Section 3.

2.4. Score-based Matching

Next, we define a matching procedure that only considers the score representation of the query $Q \in \mathcal{Q}$ resulting from the OMR step. In a first step, we convert the OMR result into a piano-roll like representation as indicated by Fig. 1d. Dealing with monophonic themes (a property that may be corrupted by the OMR step), we consider the upper pitch contour of the OMR result. Since OMR often fails at detecting the correct note durations but tends to correctly recognize the bar lines, we do not use the note durations but locally resample the pitch sequence to match the bar line constraints, see Fig. 1e. This results in a sequence of pitch values. Furthermore, since OMR often misinterprets the global clef, we convert the pitch sequence into a sequence of intervals (differences of subsequent pitches), see Fig. 1f. The interval sequence, denoted by Q^{s}, is used for the matching step. Similarly, we process a document $D \in \mathcal{D}$, this time starting with a MIDI representation. The resulting interval sequence is denoted by D^{s}.

The OMR also often fails in detecting accidentals of notes, so that a pitch may be changed by one semitone. Using the edit distance would punish a deviation of one semitone to the same extent as larger deviations. Therefore, we use a local cost measure that takes the amount of the deviations into account. For two given intervals, say $a, b \in \mathbb{N}_0$, we define the distance by

$$\delta(a, b) = \frac{\min\{|a - b|, 12\}}{12} \in [0, 1]. \tag{2}$$

In this definition, we cap the value by 12 (an octave) to be robust to extreme outliers and then normalize the value. Based on this distance, we use standard dynamic time warping (DTW) as described in [16, Chapter 4]) to obtain

$$c^{\mathrm{s}}(Q, D) := \frac{\mathrm{DTW}(Q^{\mathrm{s}}, D^{\mathrm{s}})}{|Q^{\mathrm{s}}|}. \tag{3}$$

Again we normalize by the length $|Q^{\mathrm{s}}|$. In the next section, we discuss the performance of the OCR-based and OMR-based matching procedures and show how they can be combined to further improve the results.

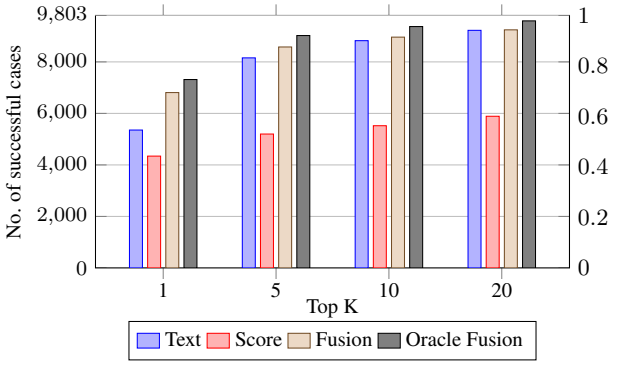

Fig. 2. Comparison of the number of top K matches for the different procedures.

3. RETRIEVAL EXPERIMENTS

We now evaluate the proposed matching procedures within a retrieval setting. In this scenario, we consider the set $\mathcal{D}$ of EDM themes as a database collection of unknown musical themes. Using a BM theme $Q \in \mathcal{Q}$ as query, the task is to identify the database document that musically corresponds to the query. Note that in this retrieval scenario there is exactly one relevant document for each query.

In our evaluation, we compare the query Q with each of the documents $D \in \mathcal{D}$ and consider the top K matches for some number $K \in \mathbb{N}$. In a search-engine-like retrieval scenario, a user typically first looks at the top match and then may also check the first five, ten or twenty matches at most. Therefore, in the following, we consider the values $K \in \{1, 5, 10, 20\}$. In the case that the top K matches contain the relevant document, we say that the retrieval process has been *successful*. Conducting the retrieval process for all 9803 queries $Q \in \mathcal{Q}$, we then count the number of successful cases. Fig. 2 shows the matching results for $K \in \{1, 5, 10, 20\}$ using four different matching procedures based on the text-based procedure from Section 2.3, the score-based procedure from Section 2.4, and two fusion procedures to be explained.

Let us start with a discussion of the text-based matching result. Considering the top match ($K = 1$), the retrieval system has been successful for 5354 of the 9803 queries, i. e., in 54.6 % of all cases. Considering the top five matches ($K = 5$), the number of successful cases increases to 8156 queries (83.2 %). This improvement can be explained by the fact that the specifications of the musical themes from the same work often differ in only a few characters, e. g. "1st Movement, 1st Theme, A" versus "2nd Movement, 1st Theme, B". Such small differences may lead to confusion among the top matches in the presence of OCR errors. Considering $K = 20$, one obtains 9225 successful cases (94.1 %), which indicates that the text-based retrieval alone already yields a good overall retrieval quality.

Next, let us have a look at the score-based matching. In the case $K = 1$, the score-based retrieval has been successful for 4342 of the 9803 queries (44.3 %). This much lower number (compared to the text-based procedure) reflects the fact that the OMR step introduces a large number of substantial errors. For example, an inspection showed that, for 1794 queries, the OMR engine was not able to produce a usable score representation. In these cases, the matching procedure was regarded as not successful. Increasing K, the results naturally improve reaching 5889 successful cases for $K = 20$ (60.1 %). To get a better picture on the overall quality of the matching proce-

Procedure	OCR	OMR	Fusion	Oracle Fusion
Mean rank	7.04	1186.26	6.24	3.34
Mean rank (capped)	3.64	9.63	2.96	2.24

Table 1. Mean ranks for the four different matching procedures. The capped mean ranks are computed by replacing the ranks above $K = 20$ to the value 21.

dures, we have also analyzed the ranking positions of the relevant documents. Recall that we obtain for each query a ranked list of the documents $D \in \mathcal{D}$, where one of these documents is considered relevant. We determine the rank of this document for each query and then compute a *mean rank* by averaging these ranks over all possible $Q \in \mathcal{Q}$. The mean ranks for all four considered matching procedures are shown in Table 1. The text-based procedure yields a mean rank of 7.04, whereas the score-based procedure results in a mean rank of 1186.26. The poor mean rank in the score-based case is the result of the unavailability of any score information for 1794 queries as mentioned above, where we set the rank to the value 4901 (half the size of $\mathcal{Q}$). Reducing the effect of outliers, we capped the rank by the value 21 (meaning that the rank is beyond $K = 20$). The mean rank of the capped values is 3.64 for the text-based and 9.63 for the score-based case. This again demonstrates that the text-based result is in average much more reliable than the score-based one.

Still, the score-based matching yields qualitatively different results than the text-based matching. We demonstrate this by fusing the matching results obtained by the two types of information. In a first experiment, we assume to have an oracle that tells us for each query which of the matching procedures performs better (in the sense that the relevant document is ranked better). The results obtained from this oracle fusion procedure yield a kind of upper limit for the joint performance of the text-based and score-based matching procedures. The results for the different values K are shown in Fig. 2, while the mean rank can be found in Table 1. For example, one obtains 7315 (74.6 %) successful cases for $K = 1$, increasing to 9592 (97.8 %) for $K = 20$. This shows that the text-based matching can be significantly improved when including the score-based information.

We now present a fusion strategy that does not exploit any oracle knowledge. The text-based matching result is taken as the basis and then refined using the score-based information. The first assumption is that the top match is particularly reliable in the case that both, the text-based and score-based matching procedures, yield the same top match. The second (weaker) assumption is that the score-based top match is somewhat reliable when it is contained in the text-based $K = 20$ top matches. The third assumption is that the score-based result is particularly reliable in the case that the cost measure defined in (3) of the score-based first (top) match is significantly lower than the cost of the subsequent second match. Based on these assumptions, we use the ranked list of the text-based matching procedure and possibly replace the top match when the condition of the second or third assumption holds whereas the conditions of the first assumption does not hold. This simple fusion strategy yields matching results as indicated by Fig. 2 and Table 1. In particular, for $K = 1$, the fusion strategy yields 6809 (69.5 %) successful cases which is close to the upper limit 7315 (74.6 %) obtained by oracle fusion.

Instead of presenting the exact details at this point, we only wanted to indicate the potential of fusing matching results. Using more refined fusion procedures could lead to results which are even closer to the upper limit indicated by oracle fusion.

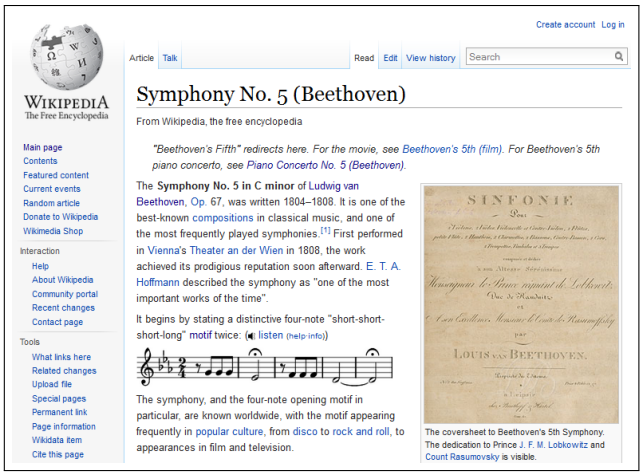

Fig. 3. Example for a typical Wikipedia website contain various types of information (text, score, image, audio).

4. APPLICATIONS AND CONCLUSIONS

In this paper, we have presented techniques for matching text-based and score-based musical information. As a case study, we used the sources from the book by Barlow and Morgenstern to serve as query input, while the EDM collection was used for evaluation purposes to serve as an example collection of digitally available musical items.

Going beyond the described (somehow controlled) scenario, we see potential of music information retrieval techniques for a much wider range of application scenarios. As mentioned in the introduction, there are millions of digitized pages of sheet music publicly available on the world wide web. Furthermore, music website as available at Wikipedia often contain information of various types including text, score, images, and audio, as shown in Fig. 3. Often the description of musical works is enriched with audio examples and score fragments of musical themes. Using similar techniques as described in this paper, one can use such structured websites to automatically derive text-based and score-based queries (and queries of other types of information such as audio or video) to look for musically related documents on the world wide web. For example, using the work specification (Beethoven, Symphony No. 5) and the score excerpt from Figure 3, one may want to retrieve sheet music representations from IMSLP or resources from less structured websites.

One main contribution of this paper was to show that matching procedures based on possibly corrupted score input (e. g., coming from OMR) may still be a valuable component, in particular within a fusion scenario where an existing classifier should be further improved.

Fusion strategies that exploit multiple types of information sources will play an important role to better cope with uncertainty and inconsistency in heterogeneous data collections, see [2]. In this context, audio-related information has been studied extensively, see, e. g., [4, 5, 17].

Future work will be concerned with integrating all available sources that describe a musical work in order to identify, retrieve, and annotate musical sources that are distributed on the world wide web.

5. REFERENCES

[1] J. Stephen Downie, "Music information retrieval," *Annual Review of Information Science and Technology (Chapter 7)*, vol. 37, pp. 295–340, 2003.

[2] 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.

[3] Nicloa Orio, "Music retrieval: A tutorial and review," *Foundation and Trends in Information Retrieval*, vol. 1, no. 1, pp. 1–90, 2006.

[4] 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 Conference on Music Information Retrieval (ISMIR)*, Paris, France, 2002.

[5] Joan Serrà, Emilia Gómez, and Perfecto Herrera, "Audio cover song identification and similarity: background, approaches, evaluation and beyond," in *Advances in Music Information Retrieval*, Z. W. Ras and A. A. Wieczorkowska, Eds., vol. 274 of *Studies in Computational Intelligence*, chapter 14, pp. 307–332. Springer, Berlin, Germany, 2010.

[6] Pierfrancesco Bellini, Ivan Bruno, and Paolo Nesi, "Assessing optical music recognition tools," *Computer Music Journal*, vol. 31, no. 1, pp. 68–93, 2007.

[7] Donald Byrd and Megan Schindele, "Prospects for improving OMR with multiple recognizers," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Victoria, Canada, 2006, pp. 41–46.

[8] Christian Fremerey, Meinard Müller, and Michael Clausen, "Towards bridging the gap between sheet music and audio," in *Knowledge representation for intelligent music processing*, Eleanor Selfridge-Field, Frans Wiering, and Geraint A. Wiggins, Eds., Dagstuhl, Germany, Jan. 2009, number 09051 in Dagstuhl Seminar Proceedings, Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany.

[9] Ana Rebelo, Ichiro Fujinaga, Filipe Paszkiewicz, Andre RS Marcal, Carlos Guedes, and Jaime S Cardoso, "Optical music recognition: state-of-the-art and open issues," *International Journal of Multimedia Information Retrieval*, vol. 1, no. 3, pp. 173–190, 2012.

[10] Christopher Raphael and Jingya Wang, "New approaches to optical music recognition.," in *ISMIR*, 2011, pp. 305–310.

[11] Harold Barlow and Sam Morgenstern, *A Dictionary of Musical Themes*, Crown Publishers, Inc., revised edition third printing edition, 1975.

[12] Ray Smith, "An overview of the tesseract ocr engine.," in *ICDAR*, 2007, vol. 7, pp. 629–633.

[13] Hervé Bitteur, "Audiveris - open music scanner," Website `https://audiveris.kenai.com`, last accessed 09/29/2014, 2013.

[14] Jacob T. Schwartz and Diana Schwartz, "The electronic dictionary of musical themes," Website `http://www.multimedialibrary.com/barlow/`, last accessed 08/07/2014, 2008.

[15] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein, et al., *Introduction to algorithms*, vol. 2, MIT press Cambridge, 2001.

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

[17] Frank Kurth and Meinard Müller, "Efficient Index-Based Audio Matching," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 382–395, Feb. 2008.