BRIDGING PRINTED MUSIC AND AUDIO THROUGH ALIGNMENT USING A MID-LEVEL SCORE REPRESENTATION

Özgür İzmirli, Gyanendra Sharma
Center for Arts and Technology
Computer Science Department
Connecticut College
{oizm, gsharma}@conncoll.edu

ABSTRACT

We present a system that utilizes a mid-level score representation for aligning printed music to its audio rendition. The mid-level representation is designed to capture an approximation to the musical events present in the printed score. It consists of a template based note detection front-end that seeks to detect notes without regard to musical duration, accidentals or the key signature. The presented method is designed for the commonly used grand staff and the approach is extendable to other types of scores. The image processing consists of page segmentation into lines followed by multiple stages that optimally orient the lines and establish a reference grid to be used in the note identification stage. Both the audio and the printed score are converted into compatible frequency representations. Alignment is performed using dynamic time warping with a specially designed distance measure. The insufficient pitch resolution due to the reductive nature of the mid-level representation is compensated by this pitch tolerant distance measure. Evaluation is carried out at the beat level using annotated scores and audio. The results demonstrate that the approach provides an efficient and practical alternative to methods that rely on symbolic MIDI-like information through OMR methods for alignment.

1. INTRODUCTION

Music can be represented in mainly three forms: audio, symbolic (such as MIDI) and printed. Historically these forms of data have remained disparate in archives and have only been associated through metadata. More recently the field of music information retrieval has been actively exploring ways to bridge the content across their different forms of existence. MIR systems dealing with large music collections depend on basic operations such as searching, matching and alignment. These operations are required to not only work with audio or MIDI formats but they should be capable of handling multi-format data including printed and hand-written scores. Finding matches and similarities across representations is of interest because these will pave the way to building integrated systems that have broad implications in research and education.

Traditional libraries contain vast collections of music on paper as well as recorded audio but lack the fine-level connection between the two formats. Incorporation of methods that connect the different representations can result in applications being more capable and multimodal. Some applications include: score retrieval by audio example; structure and harmonic analysis by audio input; transcription of performance parameters from audio superimposed onto existing scores; score following in the literal sense – following the music automatically on the printed score.

Audio is sonically rich but sound mixtures are hard to analyze and separate automatically. Symbolic data on the other hand represents music very efficiently at the note level but contains very little timbral and expressive information. The visual nature of the printed score allows musicians to read, experience and analyze music in different ways and is an indispensible part of musicians' every day experience. Each representation type has its own advantages and by connecting them in meaningful ways we can achieve greater musical understanding as well as convenient access to many forms of representation. The different kinds of information in these representations can greatly leverage our overall understanding and aid us in searching with multiple perspectives. Today, conversion between these forms presents many challenges and can be performed with varying levels of success. It is, however, easier to bridge collections in different forms through fine-grain alignment.

In this paper, we present an approach to aligning audio and printed representations of music using a mid-level score representation. We will use the term *score* to denote the sheet or printed version throughout the paper and note that it is different from the usage in score following work where it is commonly used to depict the MIDI-like symbolic sequence. The proposed mid-level representation enables us to capture sufficient pitch and score position information to guide the alignment process. Key signatures and accidentals are ignored in the recognition and therefore a tolerant distance measure that compensates for this shortcoming is proposed. In the remainder of the paper the next section outlines related and previous work. Section 3 presents the mid-level representation which is followed by a section in which a distance measure is de-

fined. We finally present an evaluation of the method on a small set of piano music and close with concluding remarks.

2. RELATED WORK

Since the introduction of optical music recognition (OMR), multiple works have been carried out in mapping and aligning different music representations, namely, the music score, audio recordings and MIDI. Multiple approaches have been employed to build state-of-the-art audio-to-score alignment algorithms. Some are based on graphical and statistical models such as the ones in [3,8,16] whereas some use the Dynamic Time Warping (DTW) algorithm to align the sequences of features extracted from both the audio and the score as in [7,12]. Work done in [13] carry out a multi-pass algorithm where they propose a method which estimates the onset times of individual notes in a post processing step to obtain an accurate audio-to-score alignment. Earlier works in audio-to-score alignment such as [14] employ the DTW algorithm and generate spectral approximations from the symbolic form in order to compute the local distance measures for the DTW.

In [8] Joder et al. propose a statistical model for music-to-symbolic score alignment where a hidden state model uses two features: chroma vectors, to model pitch content of the signal and spectral flux, to model note onsets. The approach employed in this work claims to have achieved a very precise alignment with a low complexity compared to other DTW systems. More recent work by Cont [3] discusses the use of hierarchical hidden Markov models for online and real-time audio-to-score alignment.

All these works approach the problem of alignment based on fully-notated MIDI score. To the authors' knowledge, alignment work solely based on the music sheet and its corresponding audio recordings without the use of intermediate MIDI format have not been formally used. Work on mapping, synchronization and identification of the music with audio recordings has been carried out by [5,6,11]. In [5] the authors discuss two different approaches in identifying the corresponding sections of an audio interpretation of a musical piece given the sections of the score for the same piece. The first approach where it is assumed that the performance sequence is known uses a semi-automatic approach using synchronization whereas, the second approach where the performance sequence is unknown uses matching techniques. However, OMR is used to obtain the symbolic score before employing any of the identification techniques.

Work has also been done in aligning semi-improvised music with its lead sheet [4]. This is generally more difficult as the lead sheet specifies only essential elements such as the melody, harmony, and a basic musical form. This work also stems from using the symbolic data obtained after the OMR techniques on the score.

A lot of work has been focused on solving specific problems of the OMR such as staff line detection [1] and recognizing musical symbols. Recently, in [17] the authors have employed template matching and grammatically formulated top-down models as a means of performing OMR on scanned sheet music. Since the purpose of this paper is precisely not to perform detailed OMR we refer the interested reader to two overviews of the state-of-the-art in optical music recognition [2, 18].

When audio is in the mix, chroma based representations are often used for alignment. In [9] the authors maintain that "chroma vectors drawn from representations using a logarithmic frequency scale are the most efficient features, and lead to a good precision, even with a simple alignment strategy." Here, we not only utilize a chroma based representation obtained from audio analysis but also create one from the score.

3. MID-LEVEL REPRESENTATION

In this work we restrict our method to pieces using the grand staff in which a system consists of the top staff notated with the treble clef and the lower staff with the bass clef. We have been using scanned scores from the International Music Score Library Project (IMSLP). These images are particularly challenging due to the fact that they contain skew within the page, have different print styles, their original resolutions vary and they are quite noisy. Our purpose is to perform some basic image processing operations on the digitized score and extract the relevant sections to arrive at an intermediate representation that would be useful for alignment. As a first step, prior to any image processing a binarization step is performed using Otzu's method [15] which optimizes the foreground/background classification of pixels through an exhaustive threshold search.

3.1 Overall Page Structure

The first step is to identify the overall page structure in terms of systems. A horizontal projection P_p is calculated by summing the pixel values across the page. This projection is generally a quite good representation to identify line positions in scanned scores that are reasonably straight. We assume that the original rotation of the scanned page produces a projection in which the staff lines are identifiable through local peaks. We can optionally perform an automated page rotation to correct for scanning errors using a procedure similar to the one described in the next section for individual systems. The projection P_p is then smoothed with a truncated Gaussian filter with width equal to a single staff. The position of each staff is determined by peak picking and simultaneously, the positions of the top and bottom lines of each system are determined by finding the local peaks of the unsmoothed projection in the vicinity of the peaks of the smoothed projection. Figure 1 shows part of the original score at the top and the projection resulting from that image below. The projection is aligned with the image of the two systems shown at the top. This process results in fairly reliable vertical position estimates of the systems on a single page. This segmentation is performed for all subsequent pages in the score for the piece in question.

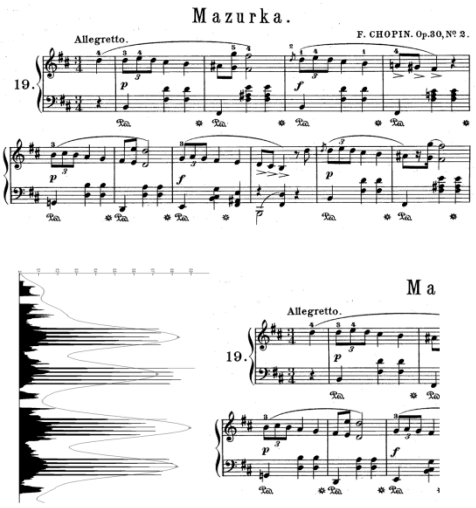

Figure 1. Top: first two systems from a scanned score. Bottom: horizontal projection and Gaussian smoothing of the top figure for locating systems in a page.

3.2 Aligning Systems and Automatic Calibration

We extract systems one by one according to their positions on a page as described above. Each system I_n, runs from C2, two ledger lines below the bottom line in the bass clef, to approximately E6 on the third ledger line above the treble clef. This image is then corrected for optimal rotation by fitting parabolas onto local peaks in the projection. The position of each line is determined by peak picking on the horizontal projection of the extracted system. We observe that the shape of the intensity distribution around each peak is correlated to how well the system is aligned – the narrowest distribution is considered the best rotation for I_n which results in maximally horizontal staff lines. We therefore, fit a least-squares parabola on the points neighboring each peak that exceeds a threshold. Since the parabolas are opening downward we find the rotation angle θ that minimizes the sum of the coefficients of the second degree terms of the parabolas for all 10 peaks. The optimal rotation angle is given by

$$\theta_n = \underset{\theta}{arg\,min}\left(\sum_i \Theta(\Psi(I_n,\theta),i)\right) \quad (1)$$

where Ψ represents the rotation of I_n by θ and Θ is the coefficient of the second degree term in the parabola equation for local peak i that serves as the relative width estimate in the horizontal projection of the rotated image.

For our purposes the rotation correction for each system turns out to be quite important. We have observed that even systems on the same page can have different rotation and skew values. In order to correctly identify notes, an adaptive reference grid delineating the note ranges is required for each system. In contrast to other approaches, our method does not remove the staff lines because the templates which are taken from actual images already include parts of those lines.

3.3 Compressing the Grand Staff

The process of finding the optimal rotation for each system also allows us to more accurately identify the positions of the lines. Next, we separate each system into two images by cutting it in half with a horizontal line that lies between the lowest line in the treble clef (E4) and the highest line in the bass clef (A3). We then merge the upper and lower halves of each system by multiplying (ORing – with pixel intensity values between 0 and 1) the content such that the positions of C4 in each part coincide. Figure 2 shows the compressed image for the first system given in Figure 1. Note that, for example, the note D4 that originally appears on the lower staff now has the correct position with respect to the upper staff. The reference grid which contains positions for the note boundaries is calculated from the distance between the top and bottom staff lines. Figure 3 shows the grid on top of a fragment of the rotated image. The regions between lines of the grid represent the C major diatonic set regardless of any accidentals that are in use.

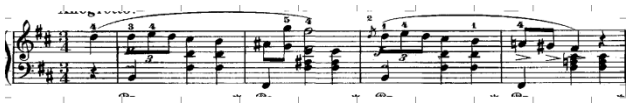

Figure 2. Compressed image of first system in Figure 1.

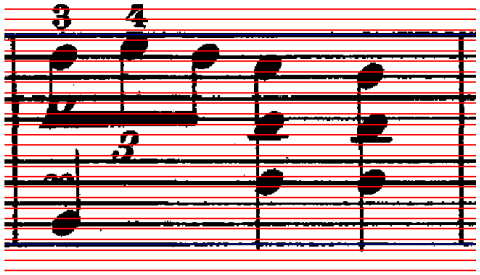

Figure 3. The reference grid for note boundaries superimposed on the optimally rotated image. The space between each pair of lines corresponds to a diatonic note.

3.4 Note Identification

The process of note identification follows a template matching approach. Three templates are constructed: one for a filled-in note head positioned between lines, one for a filled-in note head on a line and one for an empty note head. The templates are slightly larger than the oval note head and include a small portion of the surrounding lines. Their registration point is at the center of the symbol and ideally should fall either on a line or midway between two lines. The only symbols of interest are the note heads and other symbols such as stems, accidentals, rests, clefs, beams etc. are not considered. Notes are found by convolving the optimally rotated system image $\Psi(I_n,\theta_n)$ separately with each of the templates T_l. The original templates are scaled according to the line spacing of the system under consideration. The two images are represented with bipolar encoding (±1) for the convo-

lution. Local peaks indicate matches between a template and a system. We then obtain the set of all recognized notes by the union of notes recognized in all systems in the piece.

$$Q = \bigcup_{n,l} \Gamma\left[\Psi(I_n, \theta_n) * T_l\right] \quad (2)$$

Here * is the 2D convolution operator and Γ[ ] is the function for finding local peaks. Since the note recognition is done without regard to key signature or any preceding accidentals, only the notes corresponding to white keys are found. This process results in a set of recognized notes Q with 2-tuple elements $q_v=(n_v, o_v)$ each with a note index n_v that corresponds to the bins of the chromagram and an onset frame (column) number o_v. An example of the output is shown in Figure 4.

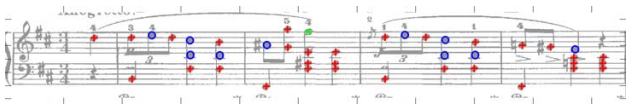

Figure 4. Recognized notes from the image in Figure 2. The original has been lightened and '+' indicates a note head on a line, 'o' between lines and 'x' an unfilled note head.

3.5 Chroma Representation from the Score

Before we define a distance function to establish a relationship between the audio and printed score representations we would like to find the most compatible frame based features that could be practically calculated from each form of the music. On the audio side we calculate an open 'audio chromagram,' A, using constant Q spectral analysis, that is not folded into one octave. The bins represent logarithmically spaced frequency ranges that are each a semitone wide and calculated with respect to a reference of A4=440Hz. We then proceed to construct a similar feature using the notes recognized from the score to form the 'score chromagram,' S. Each recognized note is placed into the chromagram in the bin representing the note and at the corresponding frame. In addition to the fundamental frequency component, the note's harmonics are also added with amplitude $1/h$, where h is the harmonic number and $h=1..H$. All components incur a fixed exponential decay to account for the passage of time, i.e. to not have the same values for the duration of the note and give more weight to the onset. The note model for a given note q_v is represented by a sequence of k-element chroma vectors

$$R^j(q_v) = (r_0^j, r_1^j, r_2^j, \ldots r_{k-1}^j)^T \quad (3)$$

$$r^j_{\varphi(q_v,h)} = \frac{1}{h} e^{-c(j-o_v)}, \quad j \geq o_v \quad (4)$$

where c is the decay rate and o_v is the frame on which the note starts. The function $\varphi(q_v,h)$ is the index of the bin in the chromagram that corresponds to note q_v and harmonic h. The resultant score chromagram is given by the summation of the note events calculated for all recognized notes:

$$S^j = E \circ \sum_{\substack{\forall q_v \in Q, \\ h=1..H}} R^j(q_v) \quad (5)$$

After the summation of note spectra, a spectral weighting is applied to the score chromagram to match its long-term spectral shape with that of the audio. The weighting E is calculated from the audio chromagram by simply averaging it across time and dividing by the maximum element. The operator ∘ denotes the elementwise multiplication of the vector E with each column of the summation that holds the unweighted score chromagram. Figure 5 shows the score chromagram for the notes of Figure 4 and the audio chromagram for the same fragment of music.

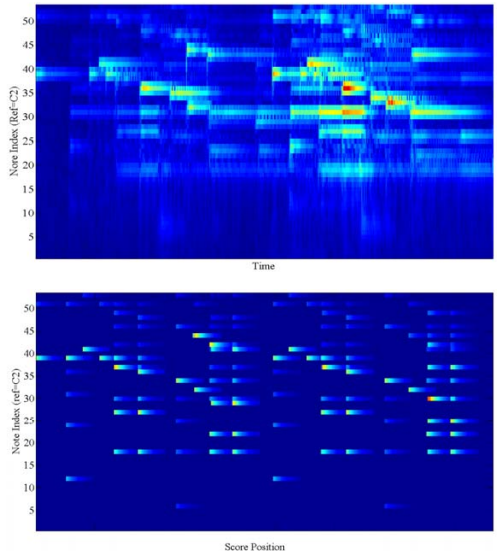

Figure 5. Top: audio chromagram. Bottom: score chromagram obtained from recognized notes as shown in Figure 4.

4. LOCAL DISTANCE AND ALIGNMENT

The defined system would have worked if the piece being analyzed was in C major and a standard distance such as a Euclidean or a cosine distance was being used. However, due to the limitation of the front-end and its notational system which is based on diatonic pitch spacing, these standard distance measures would become progressively meaningless as the keys pick up more accidentals. We therefore define a tolerant distance function between two k-element chroma vectors S (score) at frame i and A (audio) at frame j:

$$S^i = (s_0^i, s_1^i, s_2^i, \ldots s_{k-1}^i)^T, \quad A^j = (a_0^j, a_1^j, a_2^j, \ldots a_{k-1}^j)^T$$

$$b_{i,j} = \sum_{p=1}^{k-2} \frac{max\left(s_p^i a_{p-1}^j, \; s_p^i a_p^j, \; s_p^i a_{p+1}^j\right)}{\left\|S^i\right\| . \left\|A^j\right\|} \quad (6)$$

$$d_{i,j} = 1 - b_{i,j} / b_{max} \quad (7)$$

where b_{max} represents the maximum value of $b_{i,j}$.

The alignment of the score chromagram and the audio chromagram is performed using DTW. The following step size condition constrains the slope of the warping path

$$D_{i,j} = min\left(D_{i-1,j-1}, D_{i-2,j-1}, D_{i-1,j-2}\right) + d_{i,j} \quad (8)$$

Note that vertical and horizontal moves are not allowed. This ensures that the two sequences move forward at either the same frame rate or twice the other, and also that a single frame in one sequence does not map to multiple frames in the other. This preserves the monotonicity condition of the DTW for our purpose.

5. EVALUATION

We have evaluated the proposed method in various ways. Primarily the evaluation has concentrated on the accuracy of the alignment on the printed score. For this we needed the audio as well as the printed score to be annotated. The scores were taken from IMSLP's Petrucci Library which is a web site that has scanned scores for which the copyright has expired. The audio annotations for the Chopin Mazurkas were taken from The Mazurka Project (http://www.mazurka.org.uk) in which Craig Sapp collected beat-level onsets for different performances of the same piece.

We calculate the alignment accuracy with respect to two frames of reference. The first is the score where the alignment error is reported as a percentage of the staff width. The times of all beats in the audio (given by the ground truth) are mapped to score positions using the warping path produced by the DTW algorithm. The error is calculated by taking the average of the absolute differences between these numbers and the beat locations in the score given by the ground truth. The second frame of reference is the audio where the alignment error is found in seconds. The two measures are similar in nature and are not meant to provide different viewpoints, rather, they give a good sense of the average and maximum errors in the two modalities of experience: visually following the printed score while listening to the performance.

The following parameters were used for all scores and performances in the evaluation. The audio analysis was done with 50 percent overlapped windows of duration 50 milliseconds. Each column in the score chromagram represents a group of pixels in the input image. The number of pixels in a group is calculated separately for each audio file in order to make the number of score frames comparable to the audio frames. The decay rate, *c*, was determined empirically to be on the order of one beat as seen in Figure 5 but will vary from score to score depending on the density of the typesetting. Four harmonics (H) were used for the note model. The range of the note recognition was restricted to the range C2 to E6 and any notes beyond this range were ignored. The scores were scanned at 300 pixels per inch.

Table 1 shows the list of piece/score edition/performer combinations tested. The second column lists the alignment results with respect to the audio. The average absolute error and maximum error figures are given. The alignment error with respect to the score is given in the third column. The error is in pixel real distances on the digitized image. It shows the horizontal distance between the ground truth and result of the alignment as a percentage of the width of the score. It is reported as a percentage to make it independent of image resolution, however, by the same token, it could be affected by the number of measures that the publisher chose to fit in a single line. The same edition has been tested with different performers as well as different pieces from the same editor. In our tests a number of scores with heavy fonts and poor quality images did not produce acceptable alignments mainly due to the errors in the front-end. We observed that these were primarily grouped around certain publishers and that the template matching could be made more adaptive in future work to cater to even wider stylistic variations.

Piece/ Edition	Av (Max) Err. Audio (seconds)	Av (Max) Err. % score width	Performer
Mazurka 30-2 Mikuli	0.24 (1.78)	3.49 (28.16)	Mohovich
Mazurka 30-2 Mikuli	0.34 (2.07)	4.07 (19.13)	Fou
Mazurka 30-2 Mikuli	0.13 (0.84)	2.40 (13.53)	Ashkenazy
Mazurka 30-2 Klindworth	0.13 (1.27)	1.53 (11.05)	Mohovich
Mazurka 30-2 Klindworth	0.21 (2.51)	1.87 (14.31)	Fou
Mazurka 30-2 Klindworth	0.11 (0.89)	1.65 (14.19)	Ashkenazy
Mazurka 30-2 Scholtz	0.18 (1.50)	2.39 (18.41)	Mohovich
Mazurka 30-2 Scholtz	0.31 (2.17)	3.46 (19.69)	Fou
Mazurka 30-2 Scholtz	0.12 (1.14)	1.93 (17.04)	Ashkenazy
Mazurka 63-3 Mikuli	0.16 (1.97)	1.37 (10.85)	Ashkenazy
Mazurka 63-3 Joseffy	0.29 (3.41)	2.17 (23.65)	Ashkenazy
Mazurka 63-3 Kullak	0.29 (2.29)	1.99 (14.77)	Ashkenazy
Mazurka 67-1 Joseffy	0.17 (1.65)	2.22 (17.89)	Chiu
Mazurka 67-1 Klindworth	0.21 (1.70)	2.65 (18.51)	Chiu
Mazurka 68-3 Joseffy	0.36 (1.84)	4.49 (30.09)	Chiu

Table 1. Alignment errors for a number of Chopin Mazurkas by different performers and various editions of printed scores.

Results of the evaluation show that the method is able to align real-world printed scores to expressive audio performances. We have evaluated the method at the beat level to explore the possibility of more precise alignment. It can be seen from the table that the average time accuracy is quite good. We have implemented a test application that displays the score position as the music is playing based on the alignment. The tracking can be comfortably followed by eye and the application allows the viewer to see the errors as the performance unfolds. The average error figures on the score side are also good. However,

the maximum errors appear to be somewhat high. The reason for this seems to be the fact that when the last beat in a system is carried over to the next system (or a beat is aligned early from the next system) the calculated error includes the distance of the margins in between the two adjacent systems. Therefore, we do not think that these figures are as drastic as they look but appear as a delayed response while following.

Approximate matching offers many advantages to the problem at hand. With the relatively simple mid-level feature and the complexity of the recognition problem with the given less-than-ideal historical scores, recognition errors are frequent. However, the method allows for graceful recovery due to two reasons. One is the tolerant distance measure which inherently absorbs pitch errors. The other is the step condition of the DTW algorithm that prevents one sequence from stalling for extended periods. This allows for catch-up after a sequence of misdetections, rests or page segmentation errors. While selecting the templates and their detection thresholds a balance was struck between false positives and false negatives.

6. CONCLUSIONS

We have presented a mid-level score representation for aligning printed music to audio. The mid-level representation allows us to bypass sophisticated OMR techniques used for recognition and semantic analysis. The method allows for alignment through use of approximate pattern matching between the compatible features obtained from audio and score representations and therefore performs alignment within a framework in which symbolic recognition accuracy is not the primary concern. At this stage of the ongoing project, the model has been evaluated on piano music and a number of scanned scores at the beat level and the results are encouraging.

Future work will concentrate on adding sectioning and support for repeats, timbre learning from audio mixtures for more accurate note modeling, adding duration and dynamics into the note model, and catering to clef changes in the sheet. An extension of the proposed method to scores that employ systems other than the grand staff is of interest and would enable the method's application to symphonic as well as ensemble music.

7. REFERENCES

[1] Cardoso, J. S., Capela, A., Rebelo, A., Guedes, C., and Costa, J. P., "Staff Detection with Stable Paths," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 31(6), 1134–1139, 2009.

[2] Choudhury, G. S., T. DiLauro, M. Droettboom, I. Fujinaga, B. Harrington, and K. MacMillan, "Optical Music Recognition System within a Large-Scale Digitization Project," *Proc. 1st International Society for Music Information Retrieval Conference* (ISMIR), 2000.

[3] Cont, A., "A Coupled Duration-Focused Architecture for Real-Time Music-to-score Alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6), 974–987, 2010.

[4] Duan, Z., and Pardo, B., "Aligning Semi-Improvised Music with its Lead Sheet," *Proc. 12th International Society for Music Information Retrieval Conference* (ISMIR), Miami, 2011.

[5] Fremerey C., Clausen, M., Ewert S., and Muller, M., "Sheet Music-Audio Identification," *Proc. 10th International Society for Music Information Retrieval Conference* (ISMIR), Kobe, 2009.

[6] Fremerey, C., Damm, D., Muller, M., Kurth, F., and Clausen, M., "Handling Scanned Sheet Music and Audio Recordings in Digital Music Libraries," *Proc. International Conference on Acoustics NAG/DAGA*, 2009.

[7] Hu, N., Dannenberg, R. B., and Tzanetakis G., "Polyphonic Audio Matching and Alignment for Music Retrieval," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (WASPAA), New Palz, New York, 2003.

[8] Joder, C., Essid, S., and Richard, G., "An Improved Hierarchical Approach for Music-to-symbolic Score Alignment," *Proc. 11th International Society for Music Information Retrieval Conference* (ISMIR), Utrecht, 2010.

[9] Joder, C., Essid, S., and Richard, G., "A Comparative Study of Tonal Acoustic Features for a Symbolic Level Music-To-Score Alignment," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), Dallas, TX, US, March 2010.

[10] Joder, C., Essid, S., and Richard, G.,"A Conditional Random Field Framework for Robust and Scalable Audio-To-Score Matching," *IEEE Transactions on Audio, Speech and Language Processing,* 19(8), 2385 - 2397, November, 2011.

[11] Kurth, F., Muller, M., Fremerey, C., Chang, Y., and Clausen, M., "Automated Synchronization of Scanned Sheet Music with Audio Recordings," *Proc. 8th International Conference on Music Information Retrieval* (ISMIR), Vienna, 2007.

[12] Muller, M., Mattes, H., and Kurth, F., "An Efficient Multiscale Approach to Audio Synchronization," *Proceedings of the 7th International Conference on Music Information Retrieval* (ISMIR), Victoria, 2006.

[13] Niedermayer, B., and Widmer, G., "A Multi-pass Algorithm for Accurate Audio-to-score Alignment," *Proc. 11th International Society for Music Information Retrieval Conference* (ISMIR), Utrecht, 2010.

[14] Orio, N., and Schwarz D., "Alignment of Monophonic and Polyphonic Music to a Score," *Proc. International Computer Music Conference* (ICMC), Havana, Cuba, 2001.

[15] Otsu, N., "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 9, No. 1, 62-66, 1979.

[16] Raphael, C., "Aligning Music Audio with Symbolic Scores Using a Hybrid Graphical Model," *Machine Learning*, Vol. 65 (2-3), 389–409, 2006.

[17] Raphael, C., and Wang, J., "New Approaches to Optical Music Recognition," *Proc. 12th International Society for Music Information Retrieval Conference* (ISMIR), Miami, 2011.

[18] Rebelo, A., Fujinaga, I., Paszkiewicz, F., Marcal, A. R. S., Guedes, C., and Cardoso, J. S., "Optical Music Recognition: State-of-the-art and Open Issues," *International Journal of Multimedia Information Retrieval* (IJMIR), 2012.