SLAVE: A SCORE-LYRICS-AUDIO-VIDEO-EXPLORER

Verena Thomas **Christian Fremerey** **David Damm** **Michael Clausen**
Department of Computer Science III
University of Bonn, Germany
{thomas,fremerey,damm,clausen}@iai.uni-bonn.de

ABSTRACT

We introduce the music exploration system SLAVE, which is based upon previous developments of our group. SLAVE manages multimedia music collections and allows for multimodal navigation, playback, and visualization in an efficient and user-friendly manner. [1] While previously the focus of our system development has been the simultaneous exploration of digitized sheet music and audio, with SLAVE we enhance the functionalities by video and lyrics to achieve a more comprehensive music interaction. In this paper, we concentrate on two aspects. Firstly, we integrate video documents into our framework. Secondly, we introduce a graphical user interface for semi-automatic feature extraction, indexing, and synchronization of heterogeneous music collections. The output of this GUI is used by SLAVE to offer both high quality audio and video playback with time-synchronous display of digitized sheet music and content-based search.

1. INTRODUCTION

Various aspects of a piece of music can be described by different types of music documents, such as scans of sheet music, symbolic data (e.g., MIDI, MusicXML), text (e.g., lyrics, libretti, music analysis), audio recordings, and video. In modern digital music libraries large collections of these music documents are stored. The availability of digital music collections naturally leads to the necessity of providing tools to automatically process, analyze and prepare this multimedia data for an efficient and user-friendly access. Equally, user interfaces for an adequate multimodal presentation of and interaction with the music documents need to be provided. The last years have witnessed substantial progress in developing automated MIR processing procedures to compute synchronization and indexing information for multimedia music collections [1–4, 6]. In the area of digitalization of multimedia library contents, efforts both towards better automatized digitization techniques and semantic integration of multimedia documents are noticeable [5, 6]. Furthermore there are several proposals for user interfaces to access and present digital music databases in a multimodal manner [1, 6–9].

The most frequently encountered digitally available types of music representation are scanned sheet music, symbolic score data and audio recordings. Therefore most of the presented techniques and frameworks mainly focus on one or several of these data types. However, there is another type of music representation, which can provide library users, musicians and musicologists with rich information on the pieces of music. Today, most live performances are filmed and distributed to a broad audience via video DVD and television. Besides recordings of live performances also specific video productions of pieces of music are available. Hence, the extension of the functionalities of multimodal frameworks to support video documents and the development of user interfaces for video integration suggest themselves.

A holistic presentation using as many different media sources and types as possible can support the process of experiencing the music as well as analyzing the music with respect to different aspects. Prospective conductors, for example, might be interested in watching music videos to learn or compare the conducting style of different conductors. Providing tools to allow fast and smooth comparison between and browsing within interpretations are desirable for this purpose.

In this paper, we introduce the *Score-Lyrics-Audio-Video-Explorer* (SLAVE), which is based upon previous developments of our group. As enhancement, we propose the integration of videos into SLAVE to converge to a holistic exploration of music using various types of music documents in an integrated manner. Furthermore, we introduce a graphical user interface for the semi-automatic processing of multimedia music collections to generate indexing and synchronization structures as well as other derived data types. Note that for videos, we solely use the audio track to perform all required calculations.

The rest of this paper is organized as follows. The subsequent Section 2 provides information on the underlying techniques of feature extraction and music synchronization. In Section 3 the workflow for processing music collections and a GUI for a user-friendly management of the workflow are described. Section 4 presents the inter-

We gratefully acknowledge support from the German Research Foundation DFG. The work presented in this paper was supported by the PROBADO project (http://www.probado.de/, grant INST 11925/1-1) and the ARMADA project (grant CL 64/6-1).

[1] While "multimodal" refers to the perception of music through different modalities (user perspective), "multimedia" corresponds to the different media types (data perspective).

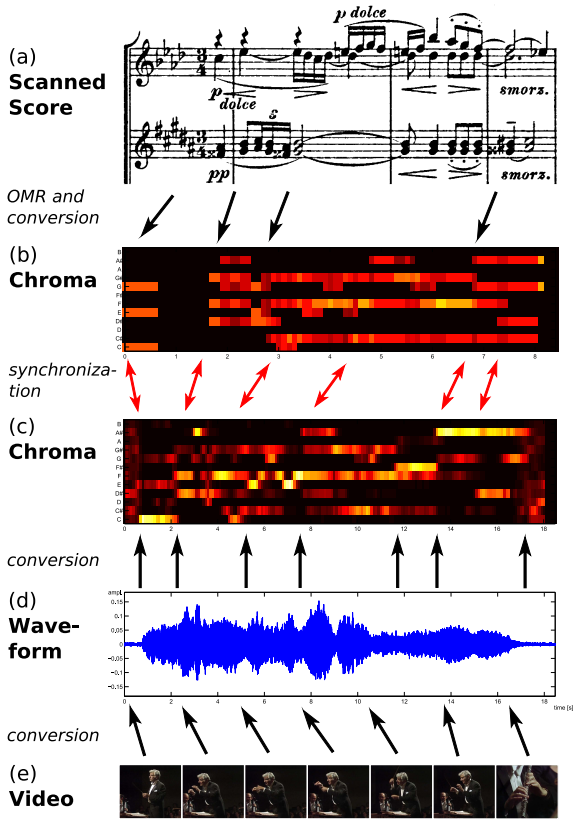

Figure 1. Illustration of the scan-video synchronization, using the first measures of the 2nd movement of Liszt's Faust Symphony. (a) Scanned sheet music. (b) Chromagram of the sheet music. (c) Audio chromagram. (d) Audio track extracted from the video. (e) Music Video. The scan-video synchronization (double-headed arrows) is obtained by chromagram-alignment, see Section 2.2.

face SLAVE and in particular the integration of video documents into this system. The paper closes in Section 5 with prospects on future work.

2. UNDERLYING TECHNIQUES

In this section, we describe the methods needed to process, match and align various types of music documents. The basic idea of the presented processing methods is to transform all music document types into a common feature representation, which allows for direct comparison and alignment independent of the input document types. In this context, chroma-based music features have proven to be a good mid-level representation [10, 11]. At first, the input documents are transformed into sequences of 12-dimensional chroma vectors, where each vector represents an energy distribution over the twelve pitch classes of the equal-tempered scale. In Western music notation, the chroma are commonly indicated by the pitch spelling attributes C, $C^{\#}$, D, . . . , B. By considering short-time statistics, these chroma features are transformed into the robust and scalable CENS features, see [12] for details. As an example, the CENS sequences for an extract of scanned sheet music and the CENS features of the corresponding video section are displayed in Figure 1 (b) and (c). Throughout this paper, chroma features with a sample rate of 10 Hz are applied, whereas CENS features of different sample rates are generated and used for alignment and indexing purposes.

2.1 Deriving Chroma-based Features

To time-align two music documents (e.g., sheet music and a video recording) describing the same piece of music, both documents are transformed into CENS features.

The transformation of scanned sheet music into CENS features requires several processing steps, see [13]. At first, using standard software for optical music recognition (OMR), the scanned score data is analyzed and transformed into musical note parameters. Subsequently, based on the gained pitch and timing information, the chroma features can essentially be computed by identifying pitches that belong to the same chroma class. The CENS sequences are gained from these features as previously described. During the feature extraction, a constant tempo of the piece of music represented by the scanned sheet music is assumed.

For a detailed description on methods for CENS feature generation of audio recordings, we refer to the literature [10, 12]. In principle, in our application the audio signal is transformed into chroma features by using short-time Fourier transforms in combination with binning strategies.

To enable the generation of CENS features from video files, the audio track of the video recording is extracted and the feature computation for audio recordings is applied.

2.2 Music Synchronization

Figure 1 gives an example of the alignment procedure for the scanned sheet music and a video recording of the first measures of the 2nd movement of Liszt's Faust Symphony. As described before, the first step for the synchronization of two music documents is, to convert both into a common and meaningful feature representation. Based on these features, multiscale dynamic time warping techniques (MsDTW) are employed to determine the synchronizations between music documents. The essential idea of MsDTW is to recursively compute alignment paths for coarse feature resolutions and project them to the next higher resolution level, where they subsequently are refined. Further details on the MsDTW method are available, e.g., in [14, 15].

During the described synchronization, we assume that the music documents match with respect to their musical structure so that only local and global tempo-variations need to be considered. In Section 3 we go into details on how to deal with structural differences during score-audio and score-video alignment. For information on synchronization of structurally differing audio recordings, see [16].

3. SEMI-AUTOMATIC DATA PROCESSING

To allow for a fast and user-friendly generation of all data files used by the SLAVE system (Section 4), the automation of all required computing steps is desirable. As a first step

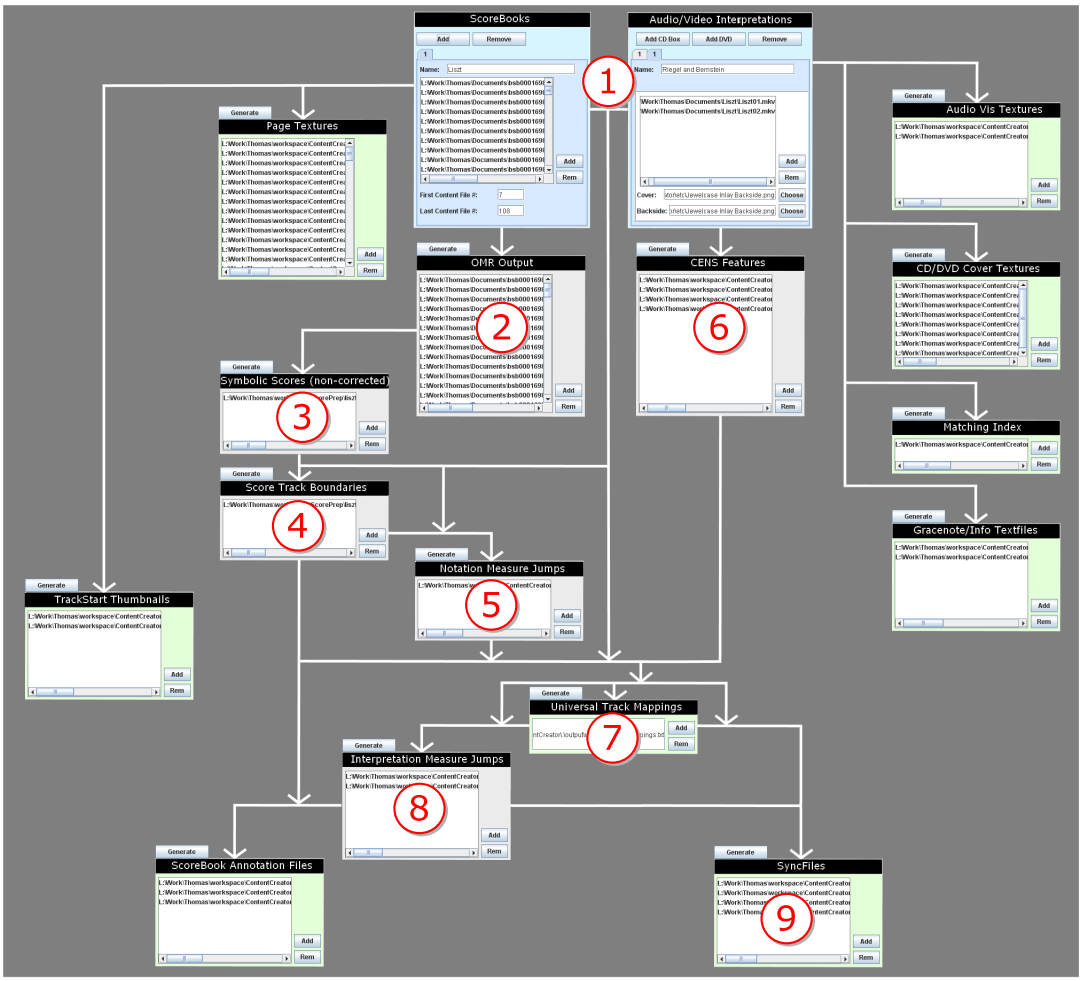

Figure 2. The *ContentCreator* interface for semi-automatic data processing.

towards full automation, we present the *ContentCreator* interface, see Figure 2, for the generation of synchronization information, indexing structures, and further data types for music collections consisting of score scans, audio recordings, and videos.

The *ContentCreator* interface aims at providing an intuitive GUI to support the process chain required to generate all metadata used by the SLAVE system. Within this context, metadata refers to data files containing information that ranges from indexing and synchronization structures to score and CD cover images used by SLAVE for visualization purposes. As shown in Figure 2, the workflow is divided into several interdependent stages. The generation of the results for each stage can be triggered individually. The integrated arrows help to clarify the dependencies between the various stages and also display the different paths for the generation of metadata. Due to the selected division of the workflow into individually triggered stages, the manual manipulation of intermediate files before continuing to the next stage is possible. At the moment, some stages merely generate default or, due to the error-prone OMR, erroneous data files. At these points, a manual rework is essential for a successful workflow. Further details on this issue are addressed in Section 3.2.

During the usage of the *ContentCreator*, three different types of data files can be distinguished. The input files (label 1 in Figure 2) mark the starting point of the process chain and currently consist of scanned pages of a music book, audio recordings, and videos. The second are the intermediate data types (labels $2-6$). These files are required for the process chain but do not contain information directly used by the SLAVE system. By contrast, the last type of files (labels $7-9$ and unlabeled boxes) contains metadata. The output stages not further mentioned in this paper generate metadata like texture data for rendering the sheet music pages, information on length and name of the audio and video tracks, and data structures for content-based retrieval.

Methods to save and restore the current state as well as the possibility to export the created metadata for runtime usage with SLAVE are provided.

3.1 Workflow for semi-automatic music alignment

In the following, we describe the chain of stages involved in the process of time-aligning a music book to various interpretations (video and audio).

As first step, the input data needs to be selected and loaded to the application (label 1). The *ContentCreator* interface enables the management of arbitrary numbers of audio and video interpretations of a piece of music (right-

hand box) and the synchronization of these interpretations to a scanned music book representing the same piece of music (left-hand box).

To extract the score information from the scanned score pages, OMR is performed (label 2) and the resulting data are subsequently merged into a single *SymbolicScore* file (label 3). This file thereby contains various music information such as note events, key signatures, time signatures, staff information, accidentals, and information on instrumentation and transposition, required for the generation of chroma-based features.

When aiming at the alignment of a whole music book to a set of video or audio recordings, the individual scanned pages of the music book need to be related to the different tracks of the music book. Therefore in the next step, a file containing information on the tracks contained in the music book is generated (label 4). Besides the musical information stored in the *SymbolicScore* files, generated in stage 3, a score also contains information on repetitions and jumps. This data is extracted from the OMR output and saved in the next stage (label 5). Together with the jump information of the video and audio interpretations, this data contributes fundamentally to the success of the synchronization process. On generating the CENS features of the scanned sheet music, the given jump information is employed to ensure structural accordance with the respective features of the video and audio recordings. Subsequently, the CENS features of all loaded audio and video files are generated as described in Section 2.1 (label 6).

Prior to the execution of the synchronization, there are two more steps to be conducted. After splitting the music book into various tracks in stage 4, these need to be mapped to the audio and video files of each loaded interpretation (label 7). Currently, manual rework is required, if the order of the videos or the audio tracks of the interpretation does not coincide with the track order of the music book. There are already proposals for a feature based, automatable creation of mapping information, which can be integrated into the *ContentCreator* interface, see [1]. Finally, the jump and repetition instructions extracted from the score scans might not be consistent with the repetitions and jumps actually performed in the specific audio or video recording. Therefore, the last stage (label 8) before the synchronization consists of the generation of this data for all loaded video and audio interpretations.

After passing all stages described before, the synchronization information for the music book and the loaded video and audio interpretations are computed (label 9), using the MsDTW approach mentioned in Section 2.2. The CENS features of the music book are computed on demand, considering the structural information of the audio or video track used for the current synchronization process.

3.2 Reworks during the workflow

The individual stages of the presented workflow are automatized as far as possible. For stages, where currently no adequate computational methods exist to enable automation, default data files, which need to be reworked by the user, are generated. The user can modify those data files or can import previously generated data files into the currently reworked stage. At the moment, parts of the *SymbolicScore* file (transposition information for the contained instruments, label 3) and the interpretation specific jump and repetition information for video and audio recordings (label 8) might need manual correction. In addition, in stage 3, 4, 5, and 7 manual adjustments might be required for complex pieces of music or low quality music book scans.

To extend the workflow managed by the *ContentCreator* to larger music collections, the applied techniques are currently integrated into the PROBADO library service system build up at the Bavarian State Library, see [1].

4. THE SLAVE SYSTEM

Recently, various computer tools were created to enable the management and presentation of multimedia music collections. However, so far those tools mostly concentrate on sheet music and audio recordings. In this section, we want to present the SLAVE system, which aims at a user-friendly and holistic exploration of music in a multimodal manner.

SLAVE is based upon the *SyncPlayer* system, presented in [7]. The *SyncPlayer* offers – besides basic audio player functionalities – the possibility of adding plug-ins for multimodal music presentation and audio analysis (e.g., a plug-in for the visualization of the musical structure of the current piece of music). SLAVE provides a renewed GUI and includes some of the techniques already available in the *SyncPlayer* as well as some new features. The new framework is envisioned as user interface for the library service system set up at the Bavarian State Library as part of the PROBADO project. First system developements towards SLAVE were recently presented in [1].

The framework consists of several user interfaces for multimodal music presentation, navigation and content-based retrieval. The central component is the *ScoreViewer* interface shown in Figure 3, which offers the visualization of the scanned pages of the underlying music book. When audio playback is started, the corresponding measures within the sheet music are highlighted based on the synchronization information created by the *ContentCreator* system described in Section 3. Some additional features of the *ScoreViewer* are automatic page turning during playback, navigation within the music book, and user-friendly music retrieval based on the query-by-example paradigm. The latter is implemented by enabling the user to select a region within the sheet music using the mouse pointer. The issued query is processed by determining the corresponding audio clipping of the currently active interpretation and performing content-based music retrieval. For details on the employed matching and indexing techniques, we refer to [17].

There might exist several interpretations of the same piece of music, which match to the given music book. The name of the currently active audio interpretation, as well as an icon showing a corresponding CD cover are displayed in the upper left corner of the *ScoreViewer* interface. To seamlessly switch to a different interpretation, a list containing information on all loaded interpretations is available by clicking on the current cover icon.

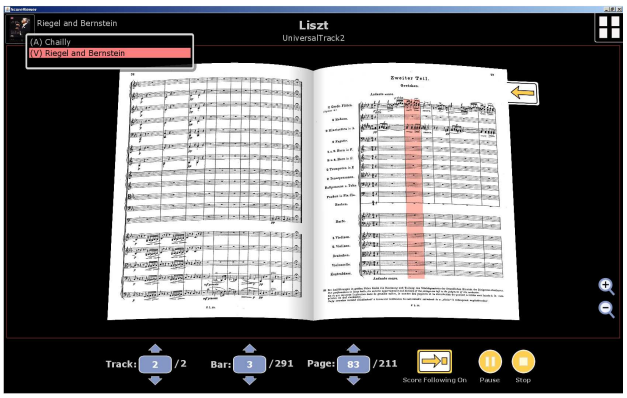

Figure 3. The *ScoreViewer* interface for multimodal music presentation and navigation. During video or audio playback the corresponding musical measures within the sheet music are highlighted. A smooth change between different interpretations of a piece of music is possible.

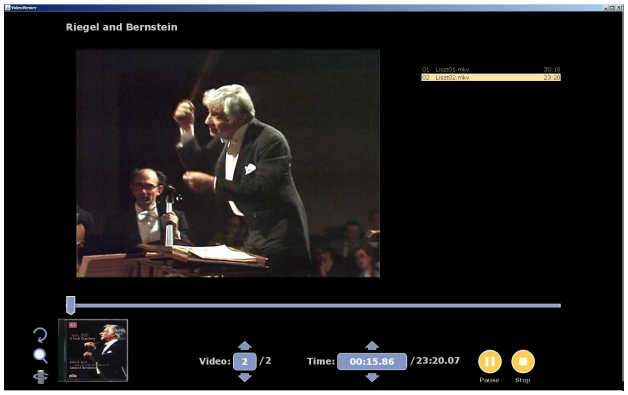

Figure 4. The *VideoViewer* shows the currently played video of the chosen video interpretation and allows browsing within the video as well as within the list of video files of the interpretation.

To include music videos, we added the *VideoViewer* interface which offers basic video player functionalities, see Figure 4. As for the audio recordings, during video playback, the corresponding measures of the sheet music are highlighted. The smooth change between different video and audio interpretations of the piece of music via the *ScoreViewer* interface enables the comparison of different music document types. Furthermore the content-based music query was extended to allow the usage of video extracts as queries and to include video interpretations in the search indices and result lists.

Although, simultaneously looking at both the *Video-* and the *ScoreViewer* might be hard for the user, having a time aligned view on the score constitutes several advantages. In longer video recordings, it might be cumbersome to search for a specific point in time within the video, whereas using the score for navigation is easier and faster. Furthermore with respect to conductings of classical music, it might be of interest to compare recordings of several different conductors at the same musical position. Using the capability of smooth switching between interpretations or performing a content-based query using the sheet music can help to facilitate these tasks.

4.1 Further Extensions

In this section we want to give a preview on further functionalities and interfaces which will be added to SLAVE for a holistic music presentation.

4.1.1 LyricsViewer

Currently, SLAVE enables the combined presentation of scanned music books, audio recordings and videos. However, text documents like libretti of operas and lyrics of song cycles are additional ingredients of digital music collections. Therefore, our current work aims at the integration of a *LyricsViewer*. Foundations for this development are the previously introduced Karaoke Display and Lyrics Seeker of our *SyncPlayer* system [18]. As for the *ScoreViewer*, the currently vocalized words of the song text will be highlighted during video or audio playback. Additionally, search mechanisms based on the lyrics will be supported by the *LyricsViewer* interface.

4.1.2 InterpretationSwitcher

In addition to SLAVE, we enhanced the *SyncPlayer* system [7] (see Figure 5), which is basically a predecessor of our new system, to support video documents.

First, the audio player component of the *SyncPlayer* was modified to allow for the playback of video files and for the inclusion of videos into playlists. Additionally we implemented the *InterpretationSwitcher* plug-in which is basically an extension of the *AudioSwitcher* plug-in [7]. The *InterpretationSwitcher* offers the possibility to switch between different audio and newly video interpretations of a piece of music during playback. On changing to a different interpretation the current playback position in the piece of music is retained and the playback seamlessly continues within the chosen interpretation.

Figure 5 shows an example of two audio interpretations and one video of the second movement of Liszt's Faust Symphony. The sliders enable the user to change to an arbitrary playback position within one of the interpretations. The playback symbols to the left of the sliders mark, which interpretation is currently playing and enable smooth switching between the interpretations. First steps towards the integration of similar functionalities to SLAVE are presented in [19].

4.2 Applications of SLAVE and the VideoViewer

The availability of video documents in the framework presented in this paper offers several advantages for musicians, musicologists, music lovers, and others. As mentioned before, prospective conductors might be able to use the proposed system to compare the work of different conductors within the same piece of music. Looking at more complex pieces of music – as orchestral works – comparing the orchestration and the arrangement of the orchestra might be of interest. Thinking of operas even stage designers, make-up artists and costume designers might have an

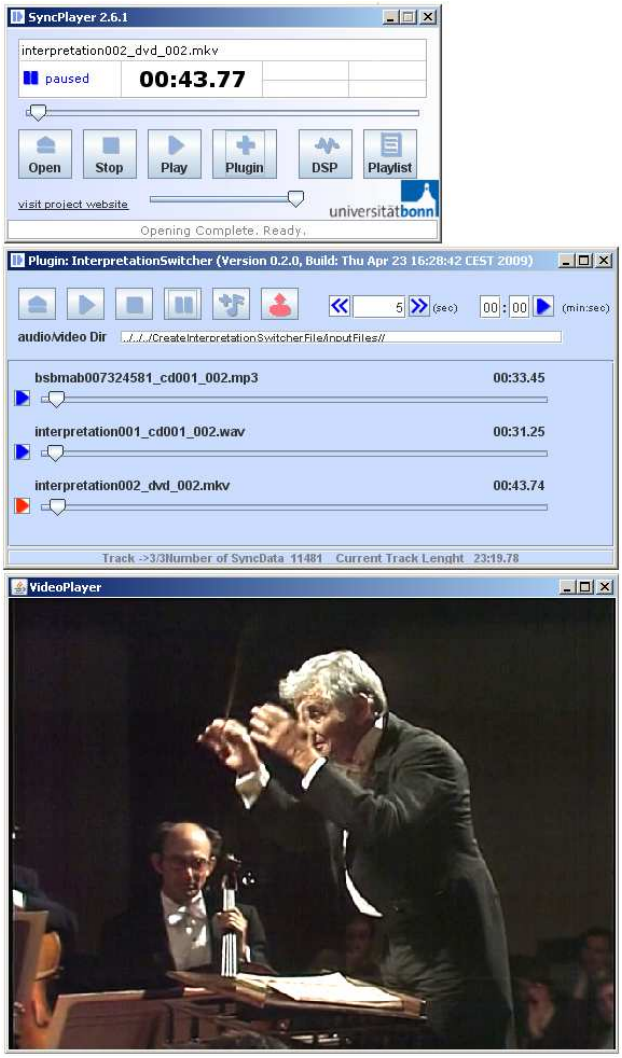

Figure 5. *SyncPlayer* with *InterpretationSwitcher* plug-in and video player. The *InterpretationSwitcher* enables the selection of several video or audio interpretations of the same piece of music. Using the sliders and the playback symbols on the left, the user can smoothly switch between and browse within the interpretations.

interest to compare different stagings. The possibility of smooth changes between various interpretations and score based navigation offers thereby significant support. Furthermore, looking at the area of dance, choreographers and dance theorists might benefit from these tools as well.

We therefore hope to experience great acceptance of the newly integrated video capabilities of our framework by the various target groups.

5. CONCLUSIONS

In this paper, we reported on a new user interface for semi-automatic processing of video, audio and score collections to generate synchronization information, indexing structures and metadata required for a holistic presentation of these heterogeneous music collections. We also presented the multimodal music management framework SLAVE and the inclusion of music videos as further music document type. Especially, we introduced the possibility of video playback and simultaneous score highlighting.

Besides the extensions described in this paper, the development of new functionalities and interfaces especially for video documents are envisioned, e.g., after the synchronization of various videos, during the playback of one reference video, the other sources can be shown time aligned to this video. For playback, only the audio track of the reference is used. This type of application will enable a more convenient and direct comparison of video interpretations.

6. REFERENCES

[1] F. Kurth, D. Damm, C. Fremerey, M. Müller, M. Clausen: "A Framework for Managing Multimodal Digitized Music Collections," *Proc. ECDL, Aarhus, Denmark*, 2008.

[2] A. D'Aguanno, G. Vercellesi: "Automatic Music Synchronization Using Partial Score Representation Based on IEEE 1599," *Journal of Multimedia*, Vol. 4, No. 1, pp. 19-24, 2009.

[3] A. Klapuri, M. Davy: *Signal Processing Methods for Music Transcription*. Springer, New York, 2006.

[4] B. Pardo: "Music Information Retrieval," *Special Issue, Commun. ACM*, Vol. 49, No. 8, pp. 28-58, 2006.

[5] BMWi: "CONTENTUS," *http://theseus-programm.de/en-us/theseus-application-scenarios/contentus/default.aspx*.

[6] C. Landone, J. Harrop, J. Reiss: "Enabling Access to Sound Archives through Integration, Enrichment and Retrieval: the EASAIER Project," *Proc. ISMIR, Vienna, Austria*, 2007.

[7] C. Fremerey, F. Kurth, M. Müller, M. Clausen: "A Demonstration of the SyncPlayer System," *Proc. ISMIR, Vienna, Austria*, 2007.

[8] A. Barat, L. A. Ludovico: "Advanced Interfaces for Music Enjoyment," *Proc. AVI, Napoli, Italy*, 2008.

[9] J. W. Dunn, D. Byrd, M. Notess, J. Riley, R. Scherle: "Variations2: Retrieving and Using Music in an Academic Settings," *Special Issue, Commun. ACM*, Vol. 49, No. 8, pp. 53-58, 2006.

[10] M. A. Bartsch, G. H. Wakefield: "Audio Thumbnailing of Popular Music using Chroma-based Representations," *IEEE Trans. on Multimedia*, Vol. 7, No. 1, pp. 96-104, 2005.

[11] N. Hu, R. Dannenberg, G. Tzanetakis: "Polyphonic Audio Matching and Alignment for Music Retrieval," *Proc. IEEE WASPAA, New Paltz, NY*, 2003.

[12] M. Müller: *Information Retrieval for Music and Motion*. Springer, Berlin, 2007.

[13] F. Kurth, M. Müller, C. Fremerey, Y. Chang, M. Clausen: "Automated Synchronization of Scanned Sheet Music with Audio Recordings," *Proc. ISMIR, Vienna, Austria*, 2007.

[14] S. Salvador, P. Chan: "FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space," *Proc. KDD Workshop on Mining Temporal and Sequential Data*, 2004.

[15] M. Müller, H. Mattes, F. Kurth: "An Efficient Multiscale Approach to Audio Synchronization," *Proc. ISMIR, Victoria, CDN*, 2006.

[16] M. Müller, S. Ewert: "Joint Structure Analysis with Applications to Music Annotation and Synchronization," *Proc. ISMIR, Philadelphia, USA*, 2008.

[17] F. Kurth, M. Müller: "Efficient Index-based Audio Matching," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 16, No. 2, pp. 382-395, 2008.

[18] M. Müller, F. Kurth, D. Damm, C. Fremerey, M. Clausen: "Lyrics-based Audio Retrieval and Multimodal Navigation in Music Collections," *Proc. ECDL, Budapest, Hungary*, 2007.

[19] D. Damm, C. Fremerey, F. Kurth, M. Müller, M. Clausen: "Multimodal Presentation and Browsing of Music," *Proc. ICMI, Chania, Greece*, 2008.