TOWARDS CROSS-MODAL SEARCH AND SYNCHRONIZATION OF MUSIC AND VIDEO STREAMS

Holger Grossmann, Anna Kruspe, Jakob Abeßer, and Hanna Lukashevich

Fraunhofer Institute for Digital Media Technology IDMT
Ilmenau, Germany
E-mail: grn@idmt.fraunhofer.de

With music markets shifting, the commercial use of music in video productions for film, TV or advertisement worldwide grows increasingly important. Our novel research project "SyncGlobal" addresses this global music licensing opportunity by developing adequate time-aware search technologies to efficiently handle sync licensing requests. The goal is to find the best acoustic or semantic matches to any video sequence from large-scale intercultural music catalogs with minimum human efforts involved. In this paper we outline conceptual issues and technology requirements derived from different application scenarios. We briefly introduce music and video segmentation, cross-modal semantic mapping strategies and time-aware music search techniques based on audio signal analysis.

Keywords: music sync search, segmentation, time-aware metadata, cross-modal semantic mapping.

INTRODUCTION AND PROJECT GOALS

The commercial use of music in video productions for film, TV or advertisement, also known as sync-licensing, is a growing market segment in the otherwise declining music business. Our current research in the SyncGlobal project is targeted towards the development of automated tools that shall support music supervisors and film producers during the creative process of assigning music excerpts to video streams. The goal is to find and synchronize the best matching music to any video sequence from large-scale intercultural music catalogs with a high level of automation.

Currently a typical work-flow involves a lot of human expertise and efforts. It starts with the segmentation of the video stream into logical units or scenes. The producer may wish to underlay some of these scenes with music in order to induce, intensify or attenuate a certain narrative element or emotional state at the recipient either in accordance with, or in contrast to, the visual modality. For this purpose producers usually describe the desired music with complex statements using some sort of vocabulary, such as "classical music, mystic, increasing intensity over time". Sometimes the producer already selects an exemplary piece of music which is familiar to her/him in order to further clarify his intentions. The requirements are specified in the so called "pitch for music" which is send out to the music labels, sometimes in combination with the original video clip, but usually without. Music supervisors at the record labels who have a deep knowledge of the respective catalogs are searching for the desired music and propose it to the producer. The returned list of tracks is then prelistened and evaluated by the producer. The decision is usually felt under

consideration of qualitative aspects and related licensing costs. The decision-making process can greatly benefit from arranging the music already along the timeline of the video before listening which is of course an expensive undertaking when not supported by automated synchronization techniques.

Our research aims at providing advanced time-aware music analysis, similarity and synch solutions that facilitate the whole work-flow while giving access to much bigger online repositories to choose the music from. Furthermore, a signal-based technology for keyword spotting from vocals tracks shall be developed in order to further extend the semantic search capabilities. Common semantic mappings between video and music shall be learned using machine learning in order to partially automate the authoring process whenever applicable. The current research builds on the results of our previous project in hybrid, adaptive music analysis, search and recommendation for global music, namely "Globalmusic2one" [3]. Representative results can be reviewed in [9] or [8].

STATE OF THE ART

Online music search today is usually limited to expert metadata or user tags, e.g. genre or mood, describing full tracks. According to their website, SynchStage [15] – a proprietary online video compositing suite for sync-licensing incorporates as well music analysis algorithms to tag the songs with relevant audible metadata on upload. The integrated authoring environment allows the user to manually offset waveforms against the video to determine the potential of a track by previewing. A similar system, however serving a different purpose was proposed by Nakano et al. [10]. DanceReProducer automatically generates dance videos by segmenting and concatenating existing dance video sequences. It employs the two types of relationships between an image sequence and music. First, in local relationships, the visual rhythms such as dance motion, camera work, etc. are synchronized with musical tempo, rhythm, and accent. Second, in content relationships, the music structure and temporal continuity of image sequence are taken into account. Doppler et al. [4] and Rubish et al. [13] present a framework for rapid score music prototyping (RapScoM) specialized on the creation of music for film and video production. The tool supports media producers by exposing a set of high-level parameters tailored to the vocabulary of films (such as mood descriptors, semantic parameters, film and music genre etc.). The semi-automated process of music production uses algorithmic composition strategies to either generate new musical material, or process exemplary material, such as audio or MIDI files.

It is natural for humans to associate the content of video and the lyrics of the accompanying music. The task of keyword spotting is well-know in speech processing [12], however it remains challenging in a case of music. Gruhne et al. [6] present an approach for phoneme recognition in popular music based on the low level acoustical features and classification techniques. A recent work of Frostel et al. [5] presents a regression model of "backness" and "height" for vowel phonemes based on common acoustic features.

APPLICATION SCENARIOS

This section sketches some of the application scenarios that are addressed within the SyncGlobal project.

Scenario 1: Video and matching music track are available, however for some reasons (i.e. licensing issues) the music need to be exchanged. The task is to find suitable music

track that shares similar characteristics in timbre, instrumentation, dynamics and structure by using the available audio track as a query.

Scenario 2: Only video is available. The task is to find an audio track that matches in both structure and appearance. Here two paradigms can be explored. First, some generic low-level rules for audio and video content are applied. For instance, when the picture in the video opens, the sound should open, become more bright, as well. Second, both video and music are segmented in homogeneous regions and these segments are associated with the metadata attributes. Based on these metadata, the optimal sound track can be found.

Scenario 3 provides a possibility to search for music by using free text keywords or predefined "catchy words". Both can be detected in the sung lyrics automatically using a novel word-spotting technology for music.

SYNCGLOBAL SYSTEM ARCHITECTURE

The architecture of the proposed audiovisual online authoring system is depicted in Figure 1. It consists of an interactive user interface, several web-services providing the system functionality and distributed databases and content repositories. The user interface implements all interactions required during the creative authoring process, starting from video segmentation and scene description over music search and browsing to music assignment and synchronization. Although the work-flow is seamlessly supported by automated functionality provided through the web-services creative user decisions based on audiovisual previews remain possible at all times. Administrative tasks, such as content ingestion, metadata enrichment or user management are handled by a separate application.

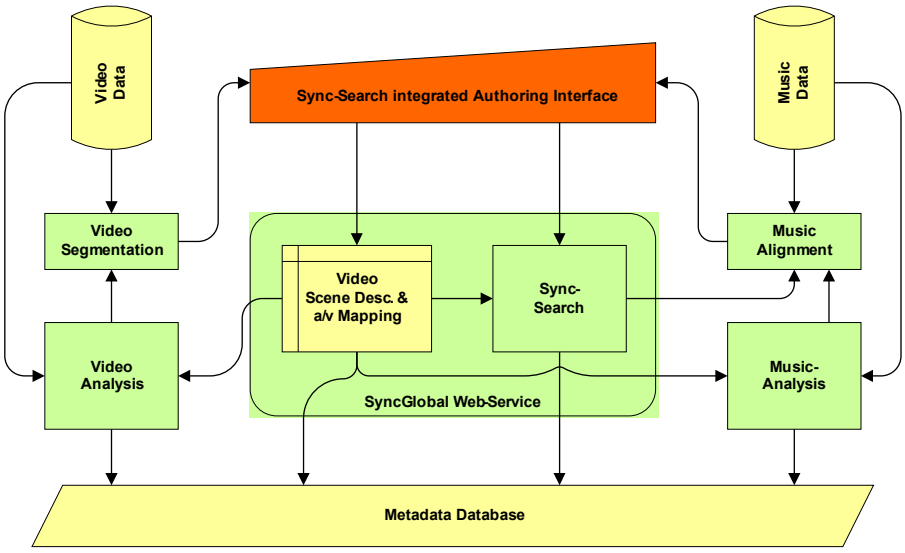

Figure 1. Architecture of the SyncGlobal system

SYNC-SEARCH PARADIGM AND CROSS-MODAL SEMANTIC MAPPING

Music and video, both are temporal streams which exhibit varying semantic qualities over time. The corresponding metadata should therefore as well be accurately assigned along the time-line of the signals in order to enable sequence-based search and usage scenarios. We are currently investigating advanced temporal analysis methods for music and

video based on regression models and dynamic texture modeling [1]. The purpose is to automatically transcribe temporal structures, gradients or progressions of musical, visual or perceptual properties as well as specific cue points within a sound file or a video clip. Rhythmic patterns are captured from music using temporal onset repetitions and from video by analyzing the dynamics of motions or cuts.

Instead of just searching for full tracks that roughly fit the requested qualities, our sync-search engine is designed to find particular segments within the tracks that are precisely in sync with the desired temporal semantics. This is essential for linking music and video in a meaningful way. Supported sync-search criteria involve among others duration, beginning or ending of a certain property, a certain shape of a property gradient or specific rhythmic patterns. In contrast to regular search expressions, temporal specifications of the search criteria are an integral part of any sync-search query.

Sync-search is deployed in our project to find music segments that match the semantics of a video scene in a certain way. However, divers mapping strategies between visual and musical characteristics seem eligible. First observations suggest that mappings are more plausible on higher levels of abstraction, e.g. rhythmic synchronicity or mood, while signal-related semantic descriptions will indeed require translations between the both modalities. Typical semantic mappings shall be further investigated during an empirical user study on the basis of TV commercials and game trailers.

MUSIC AND VIDEO SCENE SEGMENTATION

The structure of video and audio content is of particular importance for the sync-search task. It is natural when similar video content is accompanied with a similar music. In order to tackle this problem it is important to find the segment borders and the similar and/or repeating sections for both audio and video.

The task of structure analysis is well-studied for both video and music. A survey of shot boundary detection (SBD) of the research work of the annual Text Retrieval Conference Video Retrieval Evaluation (TrecVid) can be found in [14]. A less specific survey on scene segmentation and many other important issues to video analysis and video retrieval can be found in [7]. In SyncGlobal project we propose to use a two-pass algorithm for scene segmentation which uses content-based features in the first step and video production grammar rules as scene merging criteria in the second step.

An overview of state-of-the- art methods for computational music structure analysis is given in [11]. Here an audio recording is divided into temporal segments corresponding to musical parts which are further grouped into musically meaningful categories. For music segmentation in SyncGlobal project we apply a method derived from the DISTBIC algorithm successfully applied for Speaker Segmentation [2]. Firstly, the temporal positions of all possible segment changes are detected using a Bayesian Information Criterion (BIC) based novelty measure on the low-level acoustical feature vectors. Secondly, segment-class identifiers are assigned for each segment based on the BIC in an agglomerative hierarchical clustering procedure. In addition some post-processing is made restricting the minimal allowed segment length.

CONCLUSIONS AND OUTLOOK

The future of successful music marketing will depend on how well a music catalogue connects with online databases, search tools and applications. We presented several

conceptual and technical issues to be solved when developing an online music sync-licensing platform. We outlined the necessity of advanced techniques for temporal information retrieval from video and music signals in video post-production. Smart, cross-catalogue sync-search applications will not only facilitate the work of music supervisors and film producers, the almost unlimited access to distributed online catalogues is expected to further stimulate the whole music licensing business through better competition and more applicable license and pricing models.

ACKNOWLEDGEMENTS

SyncGlobal is a 2-year collaborative research project between Piranha Musik & IT AG from Berlin and Bach Technology GmbH, 4FriendsOnly AG and Fraunhofer IDMT in Ilmenau. The project is co-financed by the German Ministry of Education and Research in the frame of an SME innovation program (FKZ 01/S11007).

LITERATURE

1. *Barrington, L., Chan, B., and Lanckriet, G.* Modelling Music as a Dynamic Texture. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 602-612, Mar. 2010.

2. *Delacourt, P.* DISTBIC: A speaker-based segmentation for audio data indexing. *Speech Communication*, vol. 32, no. 1-2, pp. 111-126, Sep. 2000.

3. *Dittmar, C., Großmann, H., Cano, E., Grollmisch, S., Lukashevich, H., Abesser, J.* Songs2See and GlobalMusic2One - Two ongoing projects in Music Information Retrieval at Fraunhofer IDMT. In Proc. of the 7th International Symposium on Computer Music Modeling and Retrieval. Malaga, Spain, 2010

4. *Doppler, J., Rubisch, J., Jaksche M., and Raffaseder H.* RaPScoM: Towards composition strategies in a rapid score music prototyping. In Proc. of the Audio Mostly Conference on Interaction with Sound. Coimbra, Portugal, 2011.

5. *Frostel, H., Arzt, A., and Widmer, G.* The Vowel Worm: Real-time mapping and visualization of sung vowels in music. In Proc. of the 8th Sound and Music Computing Conference. Padova, Italy, 2011.

6. *Gruhne, M., Schmidt, K., and Dittmar, C.* Phoneme Recognition on Popular Music, In Proceedings of the 8th International Conference on Music Information Retrieval. Vienna, Austria, 2007.

7. *Hu, W., Xie, N., Li, L., Zeng, X., and Maybank, S.*. A Survey on Visual Content-Based Video Indexing and Retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. PP, no. 99, pp. 1-23, Mar 2011.

8. *Kruspe, A., Lukashevich, H., Abeßer, J., Großmann, H., and Dittmar, C.* Automatic Classification of Musical Pieces into Global Cultural Areas. In Proc. of the AES 42nd Conference on Semantic Audio. Ilmenau, Germany, 2011.

9. *Lukashevich, H., Abeßer, J., Dittmar, C., and Großmann, H.* From Multi-Labeling to Multi-Domain-Labeling: A Novel Two-Dimensional Approach to Music Genre Classification. In Proc. of the 10th Proceedings of the International Symposium on Music Information Retrieval Conference. Kobe, Japan, 2009.

10. *Nakano, T., Murofushi, S., Goto, M., and Morishima, S.* DanceReProduced: An automatic mashup music video generation system by reusing dance video clips on the Web. In Proc. of the 8th Sound and Music Computing Conference. Padova, Italy, 2011.

11. *Paulus, J., Müller, M., and Klapuri, A.* Audio-based music structure analysis. In Proc. of the International Symposium on Music Information Retrieval Conference. Utrecht, Netherlands, 2010.

12. *Rose, R.C., Paul, D.B.*. A hidden Markov model based keyword recognition system. In Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Piscataway, USA, 1990.

13. *Rubisch, J., Doppler, J., and Raffaseder, H.* RaPScoM − A framework for rapid prototyping of semantically enhanced score music. In Proc. of the 8th Sound and Music Computing Conference. Padova, Italy, 2011.

14. *Smeaton, A.F., Over, P. and Doherty, A.R.* Video shot boundary detection: Seven years of TRECVid activity. *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 411-418, 2010.

15. SynchStage – an online video compositing suite. Online at http://www.synchtank.net/synchstage