PLUNDERMATICS : REAL-TIME INTERACTIVE MEDIA SEGMENTATION FOR AUDIOVISUAL ANALYSIS, COMPOSITION AND PERFORMANCE

Dr M Grierson
Department of Computing
Goldsmiths College
New Cross
London
SE14 6NW
UK
m.grierson@gold.ac.uk
<http://www.mickgrierson.co.uk>

Abstract – This paper presents methods for real-time automated media segmentation, interactive audiovisual analysis, and media search in composition and performance tasks. In addition, we detail a use case where these tools have been deployed successfully as part of high profile public, national broadcast events, installations and exhibitions. These tools utilise a combination of data-mining and information retrieval approaches, automated sound and image segmentation techniques, and fast real-time audiovisual interaction technologies to create user-navigable media search environments, where users explore large sets of audiovisual data in fast, engaging and intuitive ways. This work is supported by the Arts and Humanities Research Council.

INTRODUCTION

Through the application of fast audio and video segmentation techniques, large amounts of sound and image data can be organised into relational databases for use in real-time applications. This paper demonstrates the successful use of these techniques for audiovisual segmentation and content description, providing users with intuitive methods for navigating large databases of audiovisual media in real-time. In addition, specific methods for displaying and organising databases of audiovisual media have been explored in a small number of high profile, user-centred use-case scenarios.

Many good methods exist for the segmentation and organisation of audio and visual media. These techniques can be divided into a number of categories, including metadata-based content organisation, symbolic segmentation approaches, and signal-based segmentation / content description. Within these three categories, a number of sub-categories exist, each with their own advantages and disadvantages, including time taken to input meta-data, and the limitations of symbolic representation techniques with respect to real sounds and images. This paper chooses to focus on purely signals-based segmentation and description methods for two main reasons. First, signals-based approaches are very effective, being deployed in a number of commercial environments, including a number of ‘music matching’ services such as Shazam. Secondly, signals-based approaches offer fully automated segmentation and content description which

allows databases to be populated with large amounts of unseen or unheard media that can then be immediately treated as source material [1].

This paper will focus on three specific signals-based techniques that form the basis of a combined toolset for deployment in interactive video scenarios: audio onset detection, motion-based shot boundary detection, and matched filtering techniques across large databases applied to both audio and video.

PREVIOUS WORK

A great deal of work has been done over the past decade on segmentation and information retrieval methods for both music and video, and a number of good review papers exist on the topic of audio onset detection [2] and video shot boundary detection [3]. In addition, there are now well-understood feature extraction and query methods for both exact [4] and approximate matching [1]. As such, research is at a level where there are working toolsets available for applying creative approaches to media information retrieval.

There is a range of work on music information retrieval-based performance systems, including BeatBox [5], Caterpillar [6], Musaicing [7], MatConcat [8], Mosievius [9], and ScrambledHackz [10], among others. However, at present there are relatively few examples of segmentation and feature extraction-based audiovisual performance systems [1, 11, 12]. One could speculate that this is because interactive video performance is not seen as an immediately useful application for segmentation and feature extraction technology.

This paper demonstrates extensions to our previous work in this area through the development of real-time prototype display systems that employ automatic segmentation, feature extraction and search as a method of audiovisual performance. Furthermore, the system here described is intended as a user-oriented video exploration system for the general public, and it is in this capacity that it has been deployed.

USING AUDIO SEARCH METHODS IN VIDEO RETRIEVAL

The ease with which digital media can be created and disseminated is resulting in an increase in the production of material. This is posing significant challenges, both in terms of technology and infrastructure, and with respect to information. There is now so much digital media that it is totally impossible to watch anything but fragments of it. The locating of sounds and images via search is facilitated by metadata, which is useful for documenting sound and image material. However, the process of creating metadata for all sound and video in a given system is becoming ever-more time consuming. This is the rationale behind automatic information retrieval approaches.

Beyond simple segmentation approaches based on shot boundary detection, video information retrieval methods tend to focus on the description and matching of specific shapes and objects in a video signal. These problems are difficult to solve. Previously, we demonstrated that information relating to some visual elements of a scene can be retrieved through signal-based analysis of the sound alone [1]. This is as a result of the close matching between sound and image in most audiovisual material, both in terms of

synchrony and semantics. This work has been demonstrated through the use of Michael Casey's C++ MAX 5 and Pure Data object, Soundspotter [1].

Soundspotter-based methods are successful with respect to feature extraction and database insertion approaches to audio search. As a real-time concatenative performance instrument, Soundspotter works by holding both the corpus and the feature database in RAM, extracting audio features from the query and finding the closest match in the database in 'real-time' (less than one FFT frame). The 89 dimensional Log Frequency Cepstral Coefficient (LFCC) features are adequate for both the recall of individual musical notes, and also timbral aspects of the sound. Furthermore, Soundspotter uses audio shingling to separate and concatenate sequences of musical and sound information.

As an exploratory test of Soundspotter's matched filter, we attempted to utilise the LFCC feature database to encode and retrieve video frames, given the assumption that, providing the video database was prepared in the appropriate fashion, periodic image features would be detectable and retrievable. The assumption proved correct, and here we demonstrate the use of Soundspotter's algorithm for video frame matching. Importantly, no visual objects are being detected: the signal is simply being matched.

Finally, this project has used the disk-based large-database version of Soundspotter (audioDB), and utilised its API to create a MAX 5 object capable of reporting a number of audio signal-based search results for each query, ranked in similarity. This is for two reasons. First, video data normally requires more memory than audio data, preventing both the corpus and database from being held in RAM. Secondly, as the intended use case is a video navigation system, more than one result was desired. Specifically, concatenation is not automated in this system. It is controlled by the user given a set of matched results.

ONSET DETECTION

We use a simple three-parameter onset detection method for segmentation of a film's audio track. More complex methods exist. However, this method was chosen due to its speed. In the current implementation, a normalised signal is tested to see if the signal is over a given threshold. If positive, the peak reference is recorded in the database. This is the start of the segment. The system then skips a given number of frames – 300ms by default, and tests to see if the signal has fallen below a separate threshold measure. If it has, the system is allowed to search for a further peak. If not, the next sample frame is tested, and if this is below the threshold, the system is allowed to continue searching.

This system is highly successful in selecting sections of speech delimited by up to 300 ms of space. Importantly, it is possible for the system to adjust its own parameters for a second pass. This is normally done if not enough segments are detected.

The onset detector has been implemented as a Max 5 Java external, seen below. The external will segment audio files of indeterminate length by loading small sections of long audio files into RAM buffers dynamically. Segments are indexed by filename and number, and accessed in the same way.

Figure 1. Screenshot of the Max 5 PeakPicker external

SHOT BOUNDARY DETECTION

Shot boundary detection is in many ways very similar to onset detection, although in this instance a slightly different approach has been taken. Video events are not normally detectable through simply searching the file for peak colour values, as this is not an adequate indication of an event in a video stream. The measure used here has been one of average frame difference.

All colour information is discarded for the purposes of analysis. A greyscale video frame (frame B) is tested against the previous video frame (frame A) in the following manner. The cellwise absolute difference between frame A and frame B is calculated, and the average of all cells in the result are used as an indication of the overall change. These values are then stored in the database alongside audio frame values.

Figure 2. Average frame difference as an indication of shot boundary

FEATURE EXTRACTION

Once all the video and audio segments have been extracted, the data is inserted segment by segment into an instance of audioDB. This can be done using the adbmax external, which interfaces with the libaudioDB shared library, created as part of the Omras 2 project. This allows for the creation of large databases that can stretch to hundreds of gigabytes. In order for video information to be inserted, it is first rasterised. Each two dimensional video image is transformed into a one-dimensional image, with the data in each row concatenated along one array, and the 8 bit greyscale values are scaled to within the range of 16bit sample values. The data can then be inserted in the same way as any audio data. Frame coherence is maintained by constraining the pixel size of the video so that it is no greater than the FFT size. This does not affect the resolution of the playback image, as it is only used for analysis. The original file is recalled once a search result is chosen.

Crucially, it is important to stress that two separate databases are used: one for audio and one for video search. At any one time, a match can be made against the audiotrack or videotrack of a current segment, providing it is already in the database.

Figure 3. A video query result using the adbmax object for Max 5

RESULTS

The three methods detailed here approach audiovisual segmentation in distinct ways. Onset detection is used for audio segmentation, shot boundary detection is used for video segmentation, and feature databases are used to measure the similarity between segments.

The type of onset detection described is successful in automatic audio segmentation, and in many circumstances is highly effective in picking out phrases that can be thought of as acoustic ‘events’. The efficacy of this approach depends on the relationship between the upper and lower peak threshold, and the size of the window. Roughly speaking, the larger the window, the larger the average segment. This approach is not without problems, however. For example, when phrases of material are indistinct or ambiguous, the phrase is less likely to be selected in one chunk. Ambiguous phrases include those where related elements are discontinuous; for example, where a sound segment contains a large pause. In these cases, long sentences can lose continuity, and musical events lose coherence. Despite this, in many cases, the onset detection method described above is successful in reducing a large dataset (such as a film soundtrack) to a large number of separate events, many of which appear coherent. This is useful for creating large databases of material for feature extraction from long sections of audio, such as film soundtracks.

The shot boundary detection method detailed above is highly successful at two separate levels. First, in the scenarios where it is currently being used, it locates static cut-style shot boundaries with close to 100 % accuracy. It is less successful in the detection of transitions, although some of these are measureable using the above technique at different thresholds. This is a known problem with this method, and some

solutions are being explored. However, it functions with a greater degree of accuracy than the onset detection method described for audio segmentation, in that almost all shot boundaries that are detected contain complete segments.

The feature extraction and database search system used in this example is the subject of ongoing research as part of the Omras 2 project. Results relating to the efficacy of this approach in audio search are well documented in [1]. However, this paper presents a novel use for this method in extraction and matching of frames from video sequences, described above. Surprisingly, results from these tests show promise. Exact video frame matches are detected with similar accuracy to exact audio frame matches. Crucially, this implementation does not attempt to search for sequences of visual material, instead searching for discrete frames, ranking the segments that contain the result frames with respect to their Euclidean distance from the query frame. In this implementation, this allows for both audio and visual sorting of segments based on a current query that can be generated either by a user, or selected from within the database itself.

In practice, this allows for real-time navigation of large audiovisual datasets containing segments of material that can be interactively sorted by the user based on features. When combined with a fast, interactive audiovisual display and playback system, the end result is a publicly accessible and easy to use video search system based on sound and image queries. In this case, the engine driving the installation is a version of the author's commercially available bespoke audiovisual composition and performance environment, *Mabuse* [12]. This system allows for the exploration of film and sound through intuitive interaction and VJ style exploration.

Study for Film and Audience was designed as a use-case for this system. It is an audiovisual installation intended to be used by the public, and attempts to engage users in a game-like fashion, encouraging them to 'play' the audiovisual segments as if in performance. Using PS3 gamepads and Wii remotes, the users can influence, control and process material taken from a selection of films and videos. The material acts as an interactive palette, and can be explored and manipulated in a number of intuitive ways based on the segmentation techniques already described. This encourages an interaction-based viewing and reviewing process, giving control of the medium over to the users.

This is also interesting for purely aesthetic and cultural reasons, as this process encourages collective refashioning of otherwise received audiovisual information, creating exploratory space for normally passive audiences to explore the process of meaning making brought about by montage and reflexive treatment. For these reasons, the piece was chosen to feature as an installation in the 2008 Sonic Arts Network EXPO festival, and then further to this, it represented the Sonic Arts category as a public installation as part of the British Academy Composer Awards in December 2008. In this way, beyond being simply a technical realisation in the context of information retrieval and interaction design, the system has been presented and accepted as an aesthetic object of high cultural value at a national level. As such, this work can be interpreted both as an artwork, and as software engineering research.

FUTURE WORK

Through continual development and integration of information retrieval processes as applied to the digital media arts, it is hoped that the work will improve both in terms of accuracy and scope. New systems are being developed and improved that allow for more complex integration of sound and image search methods for performance, and these will be integrated into new versions of the author's bespoke audiovisual composition software, Mabuse [12]. In addition, more complex approaches such as semantic speech and shape recognition could reap benefits in terms of creative applications, particularly in professional situations where large video datasets are used.

CONCLUSION

As yet, the potential impact of these technologies remains relatively unexplored when compared to their power. In addition, the utility of such approaches extends far beyond simple search. We demonstrate that these approaches can be used successfully to extend existing found-material practices, such as collage, cut-ups and plunderphonics [11]. Plundermatics, the use of semi-automated approaches in the selection, composition, production and performance of audio and visual creative work, is a new approach that is ripe for exploration.

ACKNOWLEDGEMENTS

This work is supported by the Arts and Humanities Research Council.

References

- [1] CASEY, M. and GRIERSON, M. Soundspotter / Remix-TV : fast approximate matching for audio and video performance. *In : Proceedings of the International Computer Music Conference*. Copenhagen : Denmark, 2007
- [2] COLLINS, N. A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions. *In : Proceedings of AES118 Convention*. 2005.
- [3] SMEATON, A. F., OVER, P., and DOHERTY, A. R. Video shot boundary detection : seven years of TRECVID activity. [To appear in the IEEE Computer Vision and Image Understanding.] Available at: <http://dx.doi.org/10.1016/j.cviu.2009.03.011>
- [4] WANG, A. and SMITH, J. *System and methods for recognizing sound and music signals in high noise and distortion*. United States Patent 6990453, 2006.
- [5] KAPUR, A., BENNING, M. and TZANETAKIS, G. Query-by-beat-boxing : music retrieval for the DJ. *In : Proceedings ISMIR*. Barcelona, 2004.
- [6] SCHWARZ, D. A system for data-driven concatenative sound synthesis. *In : Proceedings, DAFX-00*. Verona, Italy, 2000.
- [7] ZILS, A. and PACHET, F. Musical mosaicing. *In : Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-01)*. Limerick, Ireland, 2001.
- [8] STURM, B. MATConcat : an application for exploring concatenative sound synthesis using MATLAB. *In : Proceedings ICMC*. Miami, 2004. <http://igert.mat.ucsb.edu/pubdls/matconcat.pdf>

- [9] LAZIER, A. and COOK, P. Mosievius : feature driven interactive audio mosaicing. *In : Proceedings DAFx-03*, 2003.
- [10] <http://www.popmodernism.org/scrambledhackz/>
- [11] GRIERSON, M. *Audiovisual Composition*. Thesis, University of Kent, 2005
- [12] <http://www.mabuse.co.uk>
- [13] CASEY, M. and SLANEY, M. Song intersection by approximate nearest neighbor search. *In : Proceedings ISMIR*, 2006.
- [14] CASEY, M. and SLANEY, M. The importance of sequences in music similarity. *In : Proceedings ICASSP*. 2006.