Genetic Soundtracks: Creative Matching of Audio to Video

Jorge Gomes #, Fernando Silva # and Teresa Chambel *

#LabMAg, Faculty of Sciences, University of Lisbon, 1749-016 Lisbon, Portugal
*LaSIGE, Faculty of Sciences, University of Lisbon, 1749-016 Lisbon, Portugal

***Abstract*— The matching of the soundtrack in a movie or a video can have an enormous influence in the message being conveyed and its impact, in the sense of involvement and engagement, and ultimately in their aesthetic and entertainment qualities. Art is often associated with creativity, implying the presence of inspiration, originality and appropriateness. Evolutionary systems provides us with the novelty, showing us new and subtly different solutions in every generation, possibly stimulating the creativity of the human using the system. In this paper, we present Genetic Soundtracks, an evolutionary approach to the creative matching of audio to a video. It analyzes both media to extract features based on their content, and adopts genetic algorithms, with the purpose of truncating, combining and adjusting audio clips, to align and match them with the video scenes.**

***Index Terms*— Genetic algorithms, multimedia, entertainment, feature extraction, audio & video signal processing, video editing**

I. Introduction

The amount of home-recorded videos is dramatically increasing due to the extensive use of personal video cameras and recording tools. Without post-production, these videos typically appear very amateur and raw. One of the most common post-production processes consists on creating a soundtrack for a given video. This task, however, is usually very time consuming and daunting, because it involves manually choosing and editing the audio snippets that the user considers more appropriate for each video segment. With the increasing amount of accessible video clips and movies over the internet, the possibility of exploring alternate and creative editions in this context is an interesting subject. After all, the soundtrack can considerably affect the impact, appeal and engagement of the videos.

An evolutionary algorithm is a non-deterministic optimization method that, in each run, can generate different solutions, even if the initial parameters are the same. This way, a system that uses evolutionary algorithms can help in the creative process of defining a suitable soundtrack for a video, as the user can choose among a wide range of (apparently) viable solutions.

In this paper, we present an automated method for the matching of audio (mainly music) to videos, given an arbitrary set of audio files, and a source video selected by the user. Our method starts by automatically extracting characteristics from the video and audio files. Afterwards, the system applies a genetic algorithm in order to match the audio snippets to video, using the media characteristics to obtain the notion of fitness. We developed a tool using Java and Processing[1] to implement and test the proposed method.

Automatic video and audio content analysis is performed to extract characteristics from both media: measures through time and instants of significant changes, providing the basis for the matching process. The matching process is based on genetic algorithms (GAs), where each chromosome represents a solution – a sequence of audio snippets (or silences) with the same duration of the video. To fit the audio tracks to the video, the genetic algorithm can select, truncate, combine and discard audio clips, therefore creating a soundtrack that, at each moment, is adequate to the current video segment. This produces a pleasurable result with very little effort by the user.

The notion of soundtrack-video adequacy tries to mimic the subjective notion of fitness that users consider in manual matching. The adequacy is defined as a correspondence function between the extracted characteristics from audio and video segments, and can be customized by the users to produce a result that meets their requirements.

II. Related Work

An adequate correspondence between audio and video channels is widely recognized as an important element in the movie visualization experience. For instance, Grimes [1] conducted an empirical study where it was demonstrated that the audio-video correspondence plays an important role in attention and memory. Studies have also shown that poor sound quality degrades the perceived video image quality [2], strengthening the notion that audio and video have a strong connection in the visualization experience.

In Synesthetic Video [3], the authors explored the relation of visual and auditory properties to experience video in cross-sensorial modes, resulting in ways to hear its colors (synthesized, not matching of existing audio) and to influence its visual properties with sound and music, through user interaction or ambient influence. The motivations behind this work were accessibility, enriching users' experiences, and stimulating and supporting creativity. In [4] is performed automatic and semi-automatic selection and alignment of video segments to music. The objective of the method introduced is to create a suitable video track for a given soundtrack, which is the opposite of our work. The process is based on the detection of audio and video changes, plus camera motion and exposure, to help determine suitability between the video and audio tracks. Deterministic

[1] http://processing.org

methods are proposed for the alignment of audio and video, such as best-first search.

Evolutionary computation has been widely used in art domains, such as music generation [5] and video generation [6]. In [7], music videos are automatically generated from personal home videos, based on the extraction and matching of temporal structures of video and music, using genetic algorithms to find global optimal solutions. These solutions may involve repetitive patterns in video based on those found in the music. In MovieGene [8], the authors used genetic algorithms to explore creative editing and production of videos, with the main focus on visual and semantic properties, by defining criteria or interactively performing selections in the evolving population of video clips, which could be explored and discovered through emergent narratives and aesthetics.

III. Media Features Extraction

Our first step in matching video and audio is the automatic extraction of features from the media. The features are extracted only once for each medium, when they are first added to the system. A number of features can be extracted from both audio and video, in order to exploit synesthetic relations [3], such as audio frequency, levels or rhythm, and video color, lightness or movement. The objective is to extract from both audio and video characteristics that are perceptually relevant to the user. To demonstrate the concept, we chose to base the matching process in these two easily extractable features: video movement and audio levels. Note that the two features are intuitively related. In cinema, for example, fast moving and action-packed scenes are typically associated with a loud and vivid soundtrack, while slow moving scenes are usually accompanied by a soft soundtrack.

A. Video Analysis

To extract the video movement we use frame differencing between every adjacent frame in the video [9]. The frame differencing values are processed to obtain two metrics: i) the instants in time where scene cuts occur; and ii) the evolution of the average video movement over time. To obtain the scene cuts, each frame differencing value is compared with the average frame difference of the last N frames (correspondent to 1,5 s). If the value is greater than the average of the last by a threshold T, then it is considered a scene cut. To retrieve the average video movement through time, video frames are partitioned into blocks of 500 ms each, and the average movement inside each block of frames is calculated. An example of a video segment analysis is depicted in Figure 2 (*Video*).

B. Audio Analysis

As in the video analysis, the objective is to collect information about the average audio levels over time and to identify instants where significant changes in the music occur. The audio samples are obtained at a rate of 10 ms, and are analyzed with a succession of Fast Fourier Transforms (FFT) over time. The audio level is extracted and the frequency components of each sample are aggregated and averaged in 10 bands, one corresponding to each octave, ranging from about 21 Hz to 22 KHz. The extracted characteristics of the samples are then averaged in blocks of 500 ms, in order to obtain more meaningful data and easily processable data.

The frequency bands are used to identify significant changes in the music. These changes are captured by two methods: i) identification of significant changes in the bands values, when compared to the average values of the current music segment; ii) identification of significant changes in the bands values variation (measured by the standard deviation of the bands values). The threshold that defines what is a significant change is key in the process, and is calculated dynamically according to the following factors:

- The audio level of the sample being analyzed. Higher level values are perceptually less distinct and so the threshold increases linearly with the level.
- The time passed since the last music segment cut. It is not desirable that a music segment is too short or too long, so the threshold is higher when the sample being analyzed is too close to the last segment cut.
- Pre-defined threshold scaling factor that determines whether the algorithm should be more or less sensitive to changes in music.

This process allow us to obtain music segments that are perceptually identified by the listener. An example of two musics segmented with this method is depicted in Figure 2 (*Music 1 - Bittersweet* and *Music 2 - Fuel*).

IV. Matching Audio to Video

A. Genetic Representation and Initialisation

The matching process between audio clips and video scenes is modelled as an optimization task performed by genetic algorithms (GAs) [10]. The information regarding each 500 milliseconds audio sample is genetically encoded as a *gene*. Genes maintain extracted information such as the sample level values and sample position in the corresponding audio clip. Each string of genes is denoted as a *chromosome*, an individual if the population, which constitutes a possible solution to the matching process. Chromosomes have a pre-determined size corresponding to the number of genes necessary to keep up with the entire video. Adjacent samples belonging to the same audio snippet are placed sequentially in the chromosomes therefore maintaining its original sequence.

The initial population of chromosomes is generated through a semi-random process. Each chromosome is generated by putting together randomly chosen segments of audio clips and silences. In Algorithm 1, we summarize the method used for the creation of each chromosome in the initial population.

B. Genetic Operators and Fitness Function

As in the canonical GA, at each generation a subset of the population is selected to create a new population. Chromosomes are evaluated, selected and submitted to distinct genetic operators.

Algorithm 1 Algorithm for generating each chromosome in the initial population.

$remaining \leftarrow chromosomeSize$
while $remaining > 0$ **do**
  $option \leftarrow$ choose $audioclip$ or $silence$
  if $option = audioclip$ **then**
    $clip \leftarrow$ select one of the available musics
    $startPoint \leftarrow$ select one of the music's cuts
    $endPoint \leftarrow$ random($startPoint, musicLength$)
    $endPoint \leftarrow$ min($endPoint, chromosomeSize$)
    $snippet \leftarrow$ truncate($clip$, $startPoint$, $endPoint$)
  else
    $length <-$ random(1,$remaining$)
    $snippet <-$ silence($length$)
  end if
  add $snippet$ to chromosome
  $remaining \leftarrow remaining+$ length($snippet$)
end while

The *fitness function* evaluates the adequacy between the video and soundtracks based on the weighted composition of several factors. Weights are customizable and, therefore, the user can adjust the importance of the criteria involved. Defined criteria include:

- Correlation between audio levels at a certain time and the video movement of the contemplated scene, which should be maximised.
- Average length and number of audio snippets, which favours snippets of larger duration in order to avoid excessive changes between audio clips.
- Quantity of silence in the audio snippets, which should not be too much.
- Temporal matching between the audio sequence and the video, which tries to avoid, for instance, that an audio clip starts in the middle of a scene.

After the evaluation process, a percentage of the chromosomes are selected through *tournament selection*, to originate a new population.

Reproduction is the process by which a new set of solutions is generated, where pairs of chromosomes are combined to originate two new ones. New chromosomes are created by randomly choosing a crossover point, an instant in time that splits each of the parent audio sequences. The first part of each sequence is attached to the last part of the other sequence thus generating two new sequences. In Fig. 1, we illustrate the crossover process.

Mutations introduce changes in the chromosomes of the population. Each audio snippet in the chromosome is mutated with a given probability. If a snippet is to be mutated, mutation functions may: (i) alter the position of an audio snippet by moving it in the sequence, (ii) replace a snippet by a different one, which may or not exist in the chromosome, (iii) stretching or shrinking a certain snippet, or (iv) remove an audio snippet thus creating silence.

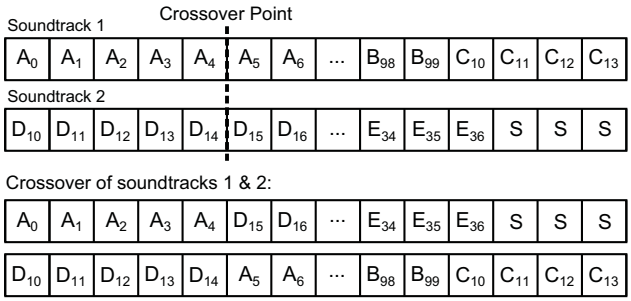

Fig. 1. Example of the crossover operator. A crossover point is chosen afterwhich the two sequences are recombined, therefore producing two new ones. X_i means audio frame i of the music X.

Decimation operator removes from the population chromosomes that violate the system's restrictions. In order to avoid radical audio changes, each audio snippet composing in a chromosome must have a minimum duration. This way, we avoid a large fragmentation of the candidate solutions and accelerate convergence to a good solution by imposing *minimal criteria*. The second restriction is related to the starting point of each audio segment. The audio segments in a chromosome should start in a transition point of the corresponding music, in order to avoid very rough transitions in the soundtrack.

V. Results

In order to validate Genetic Soundtracks, we conducted tests with a large and varied set of complete songs, and excerpts of video with distinct lengths. The measures extracted, audio level and frame differencing, are not always able to capture the intrinsic features of video scenes and audio clips. Nonetheless, the majority of the results were promising, being both compelling and pleasurable in terms of rhythm matching. Another detectable flaw was the transition between consecutive snippets. Occasionally, the transition sounds somewhat unnatural, especially when performed between snippets that are part of different audio clips. This was addressed with fade in and fade out effects, but still, the transition can sometimes feel unpleasant. A possible solution to this problem would be use of an interactive fitness, so that the user could provide the subjective evaluation missing in the automated system.

One of the main characteristics of Genetic Soundtracks is that even with the same configuration, i.e., the same video and set of audio clips, the system is capable of generating several distinct results. We will address a particular configuration both in terms of analysis and resulting matchings. The video used was a short 70 s excerpt of *Johan Hex* movie. In this excerpt, the first 44 s show a conversation between still actors, and then there is a sudden transition to an action-packed shootout, that lasts until the end of the excerpt. The audio clips used were two songs: (i) Bittersweet by Apocalyptica, a very melodic song with cellos and voice, and drums in the chorus and (ii) Fuel by Metallica, a very harsh thrash metal song. In Fig. 2, we illustrate the result of processing both songs and the video.

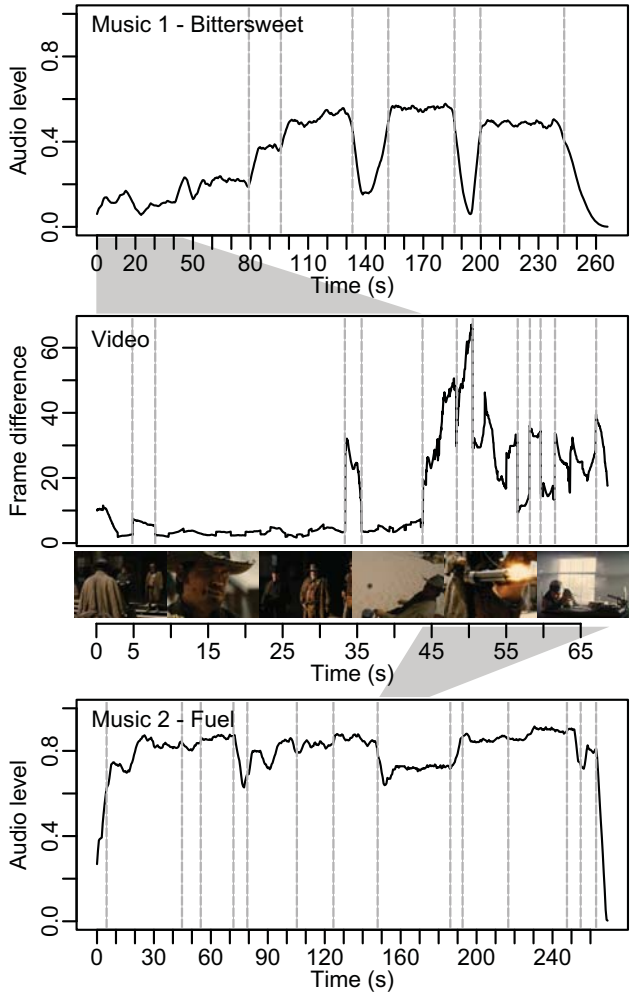

Fig. 2. The matching of two songs (Bittersweet and Fuel) and a video. The first and third graphics show the average audio levels through time, and the significant changes detected. The second graphic shows the movement of the video and scene changes detected. Gray areas represent the segment of audio selected to keep up with a given set of video scenes.

Subjectively, one of the most interesting matchings performed consisted of using the first 44s of Bittersweet followed by Fuel from 149 s to 175 s. The matching is illustrated by the gray areas between graphics in Fig. 2. The selected section of Bittersweet matches the video segment with smaller frame differencing values. The video movement then increases for which the selected portion of Fuel constitutes a good choice for keeping up with the video.

Genetic Soundtracks has also shown to be versatile enough and produced diverse matchings. For video scenes with higher frame differencing, distinct parts of the Fuel song were used. Examples of segments employed are 126 s to 143 s, 193 s to 212 s, and 106 s to 124 s, all very similar in terms of noise and rythm. Another result generated by the system consisted of using only Bittersweet for the soundtrack. For the initial video scenes, those with less movement, the system used the segment from 0 s to 50 s, with only a cello and voice. For the following scenes, the soundtrack was composed by the segment from 151 s to 171 s, which corresponds to the chorus with drums.

VI. Conclusions and Future Work

In this paper, we presented Genetic Soundtracks, an automated method for the matching of audio to a video. Genetic Soundtracks adopts an automatic method for feature extraction. The system extracts both the video movement and audio levels, which are used afterwards in the process of matching audio and video segments. We demonstrated that genetic algorithms can be applied to the truncation and combination of audio segments, even from different clips.

The immediate follow-up work will include the extension of our adequacy concept to other synesthetic relations, through the extraction of distinct media features, or the use of metadata embedded in the video and audio files. Although Genetic Soundtracks may perform automatic matching of audio to video, presenting interesting results, we intend to empower the users by involving them in the creative edition of the soundtrack. To this end, we will study innovative evolutionary combinations, to further explore the interactive definition of parameters and selection of individuals in the generations (interactive fitness) [6]. The users will therefore be able to subjectively judge different matchings of soundtracks to their videos, and guide the evolutionary search towards more personalized solutions.

Acknowledgements

This work was partially supported by LabMAg and LaSIGE through the FCT Pluriannual Funding Programme.

References

[1] T. Grimes, "Audio-video correspondence and its role in attention and memory," *Educational Technology Research and Development*, vol. 38, pp. 15–25, 1990, 10.1007/BF02298178.

[2] W. R. Neuman, "Beyond hdtv: Exploring subjective responses to very high definition television," MIT Media Lab, Tech. Rep., 1990.

[3] T. Chambel, S. Neves, C. Sousa, and R. Francisco, "Synesthetic video: hearing colors, seeing sounds," in *14th International Academic MindTrek Conference: Envisioning Future Media Environments*, ser. MindTrek '10. New York, NY, USA: ACM, 2010, pp. 130–133.

[4] J. Foote, M. Cooper, and A. Girgensohn, "Creating music videos using automatic media analysis," in *ACM Multimedia*, 2002, pp. 553–560.

[5] J. Romero and P. Machado, Eds., *The Art of Artificial Evolution: A Handbook on Evolutionary Art and Music*, ser. Natural Computing Series. Springer, 2008.

[6] T. Chambel, L. Correia, J. Manzolli, G. D. Miguel, N. A. Henriques, and N. Correia, "Creating video art with evolutionary algorithms," *Computers & Graphics*, vol. 31, no. 6, pp. 837–847, 2007.

[7] X. S. Lua, L. Lu, and H. J. Zhang, "Automatic music video generation based on temporal pattern analysis," in *12th Annual ACM International Conference on Multimedia*, ser. MULTIMEDIA '04. New York, NY, USA: ACM, 2004, pp. 472–475.

[8] N. A. C. Henriques, N. Correia, J. Manzolli, L. Correia, and T. Chambel, "Moviegene: Evolutionary video production based on genetic algorithms and cinematic properties." in *EvoWorkshops*, ser. Lecture Notes in Computer Science, vol. 3907. Springer, 2006, pp. 707–711.

[9] H. J. Zhang, A. Kankanhalli, and S. W. Smoliar, "Automatic partitioning of full-motion video," *Multimedia Syst.*, vol. 1, no. 1, pp. 10–28, Jan. 1993.

[10] M. Mitchell, *An Introduction to Genetic Algorithms*. Cambridge, MA, USA: MIT Press, 1996.