Affective Content Based Music Video Pairing System Using Real Coded GA

Chung-Hsiang Hsueh
Taiwan Evolutionary Intelligence Lab
No. 1, Sec 4, Roosevelt Road,
Taipei, Taiwan
r99921044@ntu.edu.tw

Tian-Li Yu
Taiwan Evolutionary Intelligence Lab
No. 1, Sec 4, Roosevelt Road,
Taipei, Taiwan
tianliyu@cc.ee.ntu.edu.tw

ABSTRACT

As the high quality cameras, video editing software and music composition software are more available, making videos becomes popular. However, processing home videos is time-consuming. Some popular commercial software such as iMovie and Adobe Premiere provide user-friendly interface to help people make films on their own, but adding adequate music to a clip is still a non-trivial work that highly depends on human feelings and emotions. This paper proposes an emotion-based evolutionary music video pairing system by utilizing affective information of videos and musics to remedy this problem. Empirical results show that our system is capable of pairing videos with adequate musics and self-adapting to human preferences.

Categories and Subject Descriptors

I.2.10 [**Vision and Scene Understanding**]: Video analysis

Keywords

Art and music, Computer aid design, Genetic algorithms

1. INTRODUCTION

Sometimes films made by inexperienced amateurs are not very attractive because the accompany musics do not match the videos. For experienced or professional film makers, finding appropriate clips of musics for a video may not be too difficult but is still time-consuming because they have to choose a few from numerous candidates in a huge data set. Therefore, an intelligent system that provides useful information for pairing videos with musics is desired.

2. AFFECTIVE FEATURES

Music and video are both rich in affective contents. These features and their relations with emotion are described in this section.

2.1 Audio Features

We use Marsyas system[4] to extract audio features. The descriptions of these features are shown below.

Pitch[4]: In this work, we use the *pitchextract* software in Marsyas packages to extract pitches in each time frame t, and the pitches are normalized in mel scale[3] .

GECCO'12 Companion, July 7–11, 2012, Philadelphia, PA, USA.
ACM 978-1-4503-1178-6/12/07.

The mel-scaled frequency is normalized with the maximum frequency that human can hear. The pitch feature is calculated as follows:

$$f_{pitch} = \sum_{t=0}^{T} \frac{2595log_{10}(1+\frac{f(t)}{700})}{T * 2595log_{10}(1+\frac{20000}{700})} \in [0,1]. \quad (1)$$

where $f(t)$ is the original frequency at time t, and $m(t)$ is its mel-scaled value.

Beat and tempo[2]: In this work, we use the *ibt* application in Marsyas system to extract the tempo information in each time t. The median value of these tempos is used and normalized according to the upper bound in Marsyas system.

$$f_{tempo} = \frac{mediantempo}{250} \in [0,1]. \quad (2)$$

2.2 Video Features

Film makers usually arrange emotional cues such as timing of cuts, speed of camera motion and sizing of objects with the atmosphere expressed by music. Research on the relations between visual cues and emotions indicates the components in color space[5] and the intensity of motion components[1] play important roles in conveying the emotional contents in films. Visual features we employed are as shown below.

Shot Switch Rate(SSR): HSV histogram-based detection method is used to identify shots in this work.

$$SSR = \frac{fps}{avg(shotlength)}, \quad (3)$$

where fps means frame per second, $avg(shotlength)$ is the averaged length of shots. As a result, shot switch rate represents the number of shots per second. The value of this feature is mostly in [0,1] because the average shot length is seldom smaller than fps in regular videos.

Hue, Saturation, Value: Average hue, saturation and value are the fundamental components of a picture in HSV color space.

Motion intensity (MOI): MOI is a computation of the spatial and temporal coherence according to motion vectors in frames. The value of this feature is the average length of all motion vectors in a video.

3. SYSTEM DESCRIPTION

Our system extracts low-level features of input music / video data, and utilizes real-coded GAs to map each data

onto the Arousal-Valence space. Therefore, each music and video can be represented as a point on the A-V space. The Euclidean distances between a music point and a video point on the A-V plane are used to find the pairs.

3.1 Dataset and Preprocessing

Music videos(MVs) are typical examples of music / video pairs. In this work, we use 50 popular MVs from vimeo.com and 50 popular MVs from youtube.com to perform our experiments.

3.2 Problem Encoding

In our system, each chromosome represents a mapping function that maps both video and music onto the A-V plane. The position of a video on the A-V plane is calculated below:

$$\begin{aligned} Arousal_{video} &= gene[0] * SSR + gene[1] * MOI \\ &+ gene[2] * HUE + gene[3] * SAT \\ &+ gene[4] * VAL \\ Valence_{video} &= gene[5] * SSR + gene[6] * MOI \\ &+ gene[7] * HUE + gene[8] * SAT \\ &+ gene[9] * VAL \end{aligned}$$

, while the position of a music is represented as:

$$\begin{aligned} Arousal_{audio} &= gene[10] * MedianTempo + gene[11] * Pitch \\ Valence_{audio} &= gene[12] * MedianTempo + gene[13] * Pitch. \end{aligned}$$

3.3 Fitness Function

Given several videos as queries, the fitness of a chromosome is determined by the average orders of the correct music in the candidate lists.

$$fitness = \frac{\sum_1^N (1 - \frac{order}{size(dataset)})}{N} \quad (4)$$

The maximum fitness is set to

$$MaxFitness = 1 - \frac{c}{size(database)} \quad (5)$$

, where c is a constant that indicates the worst rank of the best solution, and N indicates the number of pairs our system has to suggest.

4. EXPERIMENT

In our experiments, the chromosome length is set to 14. Crossover rate is set to 1, mutation rate is set to 0.05, α is set to 0.5, and the maximum generation is set to 200. In each generation, random choices of $\frac{1}{10}$ videos from the dataset are used to evaluate the chromosomes. Their fitnesses are determined by Function 4. The maximal fitness is set to 80% so that the correct music should be listed in the top 20 percent of the candidate list suggested by best chromosomes. The results are compared and discussed in the next section.

4.1 Validate with Testing Data

In this section, we use vimeo MVs to train our system, and use Youtube MVs to test its performance. In this trial, the parameters learned from GA is also used to map the testing data. Each chromosome is applied to pair 10 percent videos from the testing data as well. Figure 1 depicts the fitness growth averaged from 10 trials. The results show that our system is able to find correct pairs in testing data with the aid of affective mapping learned from training data.

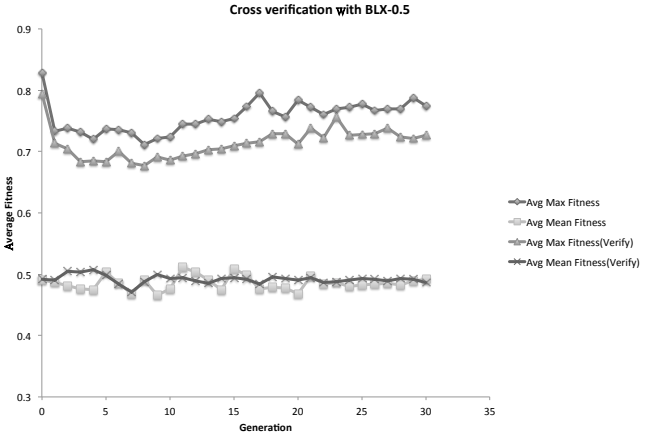

Figure 1: Cross validation with Vimeo and Youtube MVs.

5. CONCLUSION

In this work, a evolutionary framework for music video pairing is proposed. The validation with testing data demonstrates that the proposed framework is applicable to extract the implicit knowledge for pairing musics and videos. Currently, the real-coded GA offers some information of feature importance, but the information is not quantitative. In order to learn more knowledge from feature importance analysis, the value of genes should be limited. In the future, subject test evaluating pairs of home-made videos and pre-included soundtracks in some non-linear editors such as iMovie and Youtube on-line editor should be implemented to validate this system. As shown in our empirical results, combining evolutionary computations and affective models provides a good start for a smarter video editing solution.

6. REFERENCES

[1] A. Hanjalic. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7:143–154, 2005.

[2] J. Oliveira, F. Gouyon, L. Martins, and L. Reis. IBT: A Real-time Tempo and Beat Tracking System. In *Proc. Int. Conf. on Music Information Retrieval*, 2010.

[3] E. Stevens, Stanley Smith; Volkman; John; & Newman. A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, 8:185–190, 1937.

[4] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *Speech and Audio Processing, IEEE transactions on*, 10(5):293–302, July 2002.

[5] P. Valdez and A. Mehrabian. Effects of color on emotions. *Journal of Experimental Psycology*, 123:394–409, 1994.