Multi-Modal Analysis of Movies for Rhythm Extraction

Devon Bates and **Arnav Jhala**
Computational Cinematics Studio
University of California Santa Cruz
{dlbates, jhala}@soe.ucsc.edu

Abstract

This paper motivates a multi-modal approach for analysis of aesthetic elements of films through integration of visual and auditory features. Prior work in characterizing aesthetic elements of film has predominantly focused on visual features. We present comparison of analysis from multiple modalities in a *rhythm extraction* task. For detection of important events based on a model of *rhythm/tempo* we compare analysis of visual features and auditory features. We introduce an audio tempo function that characterizes the pacing of a video segment. We compare this function with its visual pace counterpart. Preliminary results indicate that an integrated approach could reveal more semantic and aesthetic information from digital media. With combined information from the two signals, tasks such as automatic identification of important narrative events, can enable deeper analysis of large-scale video corpora.

Introduction

Pace or tempo is an aspect of performance art that is aesthetically important. In film, pace is created through deliberate decisions by the director and editors to control the ebb and flow of action in order to manipulate the audience perception of time and narrative engagement. Unlike the more literal definition regarding physical motion, pace in film is a more intangible quantity that is difficult to measure. Some media forms, like music, have pace as an easily identifiable feature. Music has a beat which along with other aspects make pace readily recognizable. Modeling pace is important for several areas of film analysis. Pace is shown to be a viable semantic stylistic indicator of genre that can be used to detect a film's genre. The pace can also be useful for scene and story section detection. Pace adds a level of film analysis that adds more possible areas of indexing and classifying films. Film as a medium is unique from music in that it is a combination of audio and visual components. While, we tend to focus consciously on the visual film components, there is significant information in the audio signal that is often overlooked in analysis. While the video components of pace are important in contributing to the audience perception of pace, the audio components are also very important for pace, particularly in signifying narrative importance of events.

The visual pace of a film is determined by the content and movement of a shot as well as the length of a shot. In other words, the tempo is higher for shots that are shorter. That effect can be especially seen in movies by Michael Bay who has developed a style that involves rapid cuts between shots. That style has a fast tempo and is more often used in action movies. Another aspect of a shot that contributes to pace in a film is the amount of motion in the frame during the duration of the shot. To illustrate the effect of motion on the pace or tempo of the shot, we can imagine the same scene of two people talking but in one take one of the people is pacing nervously while they speak and in the other the two characters are stationary. If the two takes are the same in every way except that, the first take would be perceived to have a faster tempo because of the additional motion in the frame. In terms of automatic tempo detection, visual pace is a useful measure. The main factors of visual pace are shot length and motion energy. These can be calculated automatically, so pace can be generated without manual annotations.

The visual pace function presented by Adams et. al (Adams, Dorai, and Venkatesh 2002) is a useful tool in the automatic detection of the visual pace of a film. The function uses the shot length s in frames and the motion m of each shot n. The pace function also calculates the mean μ and standard deviation σ of the motion and shot length throughout the entire movie and uses those to weight the motion and shot length for individual shots. The visual pace function is defined as :

$$P(n) = \frac{\alpha(med_s - s(n))}{\sigma_s} + \frac{\beta(m(n) - \mu_m)}{\sigma_m}$$

Auditory Pace Function

We define audio pace similar to visual pace with changes in audio characteristics between shots to calculate pace features. The pace function we define, utilizes time domain characteristic — loudness, and a frequency domain feature — brightness, to characterize the audio signal. The pace is then calculated with a similar pace function that compares the changing characteristics. We use $L(n)$ to represent the average loudness and $B(n)$ to represent the brightness of each shot n. The medians med_L and med_B represent the medians of the brightness and loudness over all shots in the film. The standard deviations of the changes are also used.

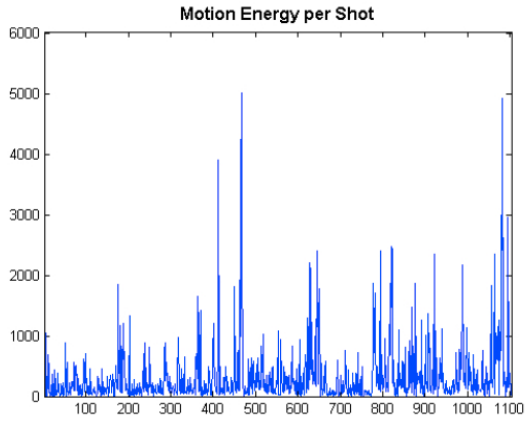

Figure 1: Motion energy per shot The average motion energy and value of start frames are all that is needed to calculate the pace function, P(n)

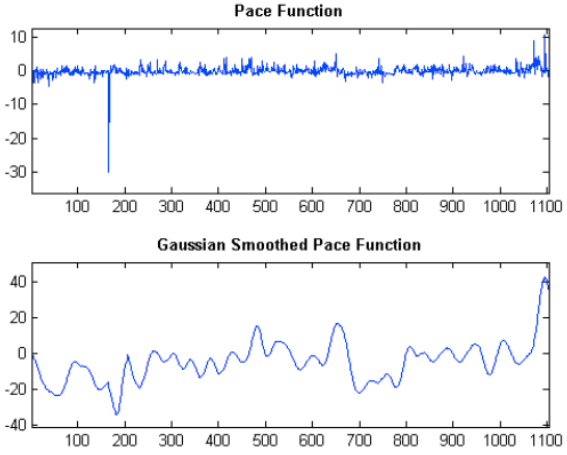

Figure 2: Pace function with Gaussian smoothing

The audio pace function if structured thus:

$$P_{audio}(n) = \frac{\alpha(B(n) - med_B))}{\sigma_B} + \frac{\beta(L_n - med(L))}{\sigma_L}$$

Loudness in an audio signal means the average energy of the signal through the duration of the shot in decibel form. Higher levels of loudness correspond to higher pace. The brightness is the centroid of spectral frequencies in the frequency domain. The higher the brightness of the signal, we can assume there are more components of the audio and that would indicate higher pace. There are many different characteristics or measurements that could be used to characterize an audio signal. In this paper, we chose a simple model with two features that corresponded well to the visual analogue.

Audio and Video Processing : We analyzed two film clips that were 25 frames/second with frame sizes of 640px by 272 px. In all, that constituted 157,500 frames for *Scream* and 136,500 frames for *Cabin in the Woods*. We imported and processed the frames in sections of 10,000

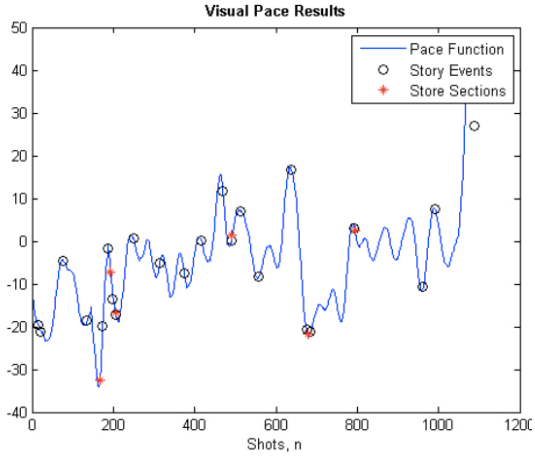

Figure 3: Visual pace results

frames and compiled the results. After importing these frames, the first step was to perform *cut detection* to detect the start frames for each shot in the film. The d_c image is the average of n by n squares of the original image and then the difference is calculated from frame to frame. The size of the window must be big enough to eliminate the noise. We used histograms of the d_c image with 4 bins and performed χ^2 difference between each frame to detect the magnitude of the change in each frame. The number of histograms bins have to be high enough to detect changes but not too high that they over detect change in noise. The average motion energy of the frame is also calculated by simply calculating the difference between n by n squares in the d_c image. So if a shot is detected from frames f_1 to f_2, the average motion energy of that shot is calculated by finding the sum of differences between dc image squares between f_1 and f_2, divided by the number of frames f_1-f_2. The results of the pace function are then analyzed to find the edges that have steep slopes. There are two levels of edges detected, the smaller is used to detect story events and the larger indicates story sections. We present a breakdown of the story sections by the range of frame numbers for Scream along with a description of the events during that story section below in Table 1.

The story events detected using the video components of the film are a separate list from the story events detected using the audio components. It is assumed that some events will be detected by both methods and some will only be detected by one. To account for that we combined story events that occur within the length of the frame used for non-maximum suppression, in this case 10 shots. We show details of our results in a section of frames where we have manually annotated story events to correspond to the timestamps extracted from analysis of the two modes. These results are for the story section from shot 792 to shot 1081. This story section contains a number of typical horror elements. It is the final set-piece and showdown between Sydney and the murderer Ghostface. It starts when he reveals himself by stabbing Billy and ends with the reporter Gale

Frame Range	Description
1- 29939	Murder of first girl and aftermath
29940-43112	Ghostface attacks Sydney
43113-44875	Billy shows up and Sydney suspects that he's Ghostface
44876-79772	Billy gets off the hook and fights with Sydney
79773-102258	All the teens go to the party, where Sydney and Billy makeup
102259-116637	Ghostface kills Tatum and Billy
116638-146020	Sydney tries to get away and finds out the truth
146021-150000	Aftermath of the killings are resolved

Table 1: Story sections from *Scream*

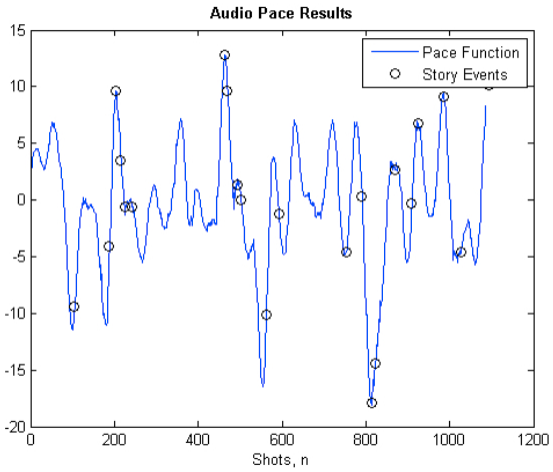

Figure 4: Audio pace results

Weathers killing Billy to end the rampage.

Figure 5 and classification results from Table 2 show comparison of the audio pace and visual pace separately. There are some sections where the two functions overlap, but in other areas they are mirror edges. From shots 1 to 200 the functions are almost mirror images, but they align almost perfectly from shots 850 to 1000. Each method has strengths and weaknesses, however neither can paint a sufficiently vivid picture of the film without the other. Overall, this preliminary analysis indicates that audio pace wasn't an appropriate way to detect story sections but was useful in detecting story events. Similarly, visual pace was a successful way to detect story events. However, during periods of sustained visual action, there was some difficulty in detecting a full set of events. In the horror genre, there are often periods of sustained action making visual pace insufficient to detect a full range of story events.

Pace is an important feature to be able to detect story structure within a film. Our techniques were limited in detecting the structure of the film without any underlying semantic meaning. Pace can detect events, but not necessarily the tone of those events. It is easy to see how the basic technique of finding pace can be expanded into not only detecting story events and sections but also characterizing them which would be an invaluable tool for automatic video annotation.

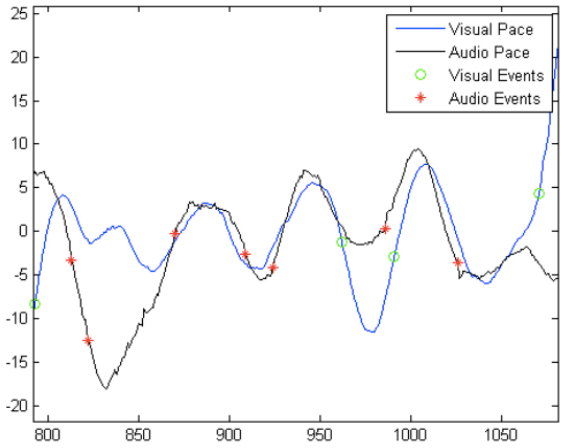

Figure 5: Pace results for both methods

Shot	Event Description	A or V
793	Ghostface "stabs" Billy	A
822	Sydney falls off the roof	A
870	Gale realizes her camera man is dead	A
909	Sydney locks Randy and Stu out of the house	A
924	Billy shoots Randy	A
962	Billy tells Sydney that he killed her mom	V
991	Billy stabs Stu to cover up their crimes	A/V
1026	The reporter holds up the bad guys with a gun	A
1071	Sydney kills Stu by dumping a TV on him	V
1089	Gale shoots Billy	V

Table 2: Story events from a section of *Scream* results show how both the audio and visual pace components can miss valid events.

Discussion

This is a preliminary study to compare models of aesthetic elements in film from diverse techniques in visual and audio processing. This is worthwhile due to the richness and diversity of information that could be gained from integra-

tion of information from these modalities. Audio processing could be better integrated with video due to the similarity of underlying signal processing and modeling techniques but there is much work in that area that needs to be done for any semantic and narrative-based analysis of audio. In this work, we focused on low-level features of audio. Much work in the audio analysis community has been done on deeper analysis of audio with respect to semantic elements such as narrative modules, tone, scene transitions, etc. There are also features that could be utilized to characterize genre, character, and other classes. Results in this paper motivate more work on integrated multi-modal analysis. Future work on integrating textual medium for such analysis in addition to video and audio could prove promising. A possible area future work would include analysis to see if these results are applicable to other genres. This paper used a viable audio pace function using loudness and brightness, but future work could also include investigating different audio features to improve models of pace that utilizes these audio features. Loudness and brightness work well in the horror genre, but other features might be required to characterize pace in genres such as romance, comedy etc. This paper also shows that there is information in both the audio and visual signals that can be mapped to the narrative structure of films. This has implication for generative content, and creating audio signals that match a preexisting narrative structure or even an existing video signal.

References

Adams, B.; Dorai, C.; and Venkatesh, S. 2002. Toward automatic extraction of expressive elements from motion pictures: Tempo. *Multimedia, IEEE Transactions on* 4(4):472–481.