

NOVEL APPROACH TO DETERMINING TEMPO AND DRAMATIC STORY SECTIONS IN MOTION PICTURES

Brett Adams[†], Chitra Dorai[‡], Svetha Venkatesh[†]

Department of Computer Science[†]
Curtin University of Technology
GPO Box U1987, Perth, 6845, W. Australia
{adamsb, svetha}@cs.curtin.edu.au

IBM T. J. Watson Research Center[‡]
P.O. Box 704, Yorktown Heights
New York 10598, USA
dorai@watson.ibm.com

ABSTRACT

This paper presents an original computational approach to extraction of movie tempo for deriving story sections and events that convey high level semantics of stories portrayed in motion pictures, thus enabling better video annotation and interpretation systems. This approach, inspired by the existing cinematic conventions known as film grammar, uses the attributes of motion and shot length to define and compute a novel continuous measure of *tempo* of a movie. Tempo flow plots are derived for several full-length motion pictures and edge detection is performed to extract dramatic story sections and events occurring in the movie, underlined by their unique tempo. The results confirm reliable detection of actual distinct tempo changes and serve as useful index into the dramatic development and narration of the story in motion pictures.

1. INTRODUCTION

Increased use of the rich digital video medium on the Internet and corporate intranet archives for applications ranging from broadcasting and product demonstrations to on-demand, online education and training has highlighted the **inadequacy** of current tools for automated video content understanding and indexing. While most research has sought solutions to the problem along the lines of simple extensions of interrogative techniques of the relatively mature textual and image query research, it has become apparent that these fall short in mining meaning from the unique modes open to the video medium. High-level semantic annotations of the video content are very crucial to avoid laborious search over huge collections of frame-oriented low-level features, and to seamlessly transform natural human descriptions of the content to automatically computable entities.

We describe a novel approach, inspired by the existing film grammar [1], [2, p. 119], [3, p. 189], to computationally determine the expressive elements of motion pictures conveyed by the manipulation of lighting, color, camera

movements, editing, etc., for high level video interpretation and reasoning. Our project, guided by this grammar, concentrates on the extraction of high-level semantics associated with the expressive elements and the form of story narration in movies. It differs from many recent approaches in that while others have sought to model specific events occurring in a particular domain, our research tries to understand the “expressiveness” of the film medium and the thematic units (high-paced sections, tranquil events, etc.) underscored by the expressions, that are pervasive regardless of the genre or domain of the story.

One expressive element, often discussed in film appreciation is *pace* or *tempo* that gives a sense of a story’s *experienced time*. Tempo is defined by [4] as being influenced “in three ways: by the actual speed and rhythm of movement and cuts within the film, by the accompanying music, and by the content of the story”. Sobchack says that “[Tempo] is usually created chiefly by the rhythm of editing and by the pace of motion within the frame” ([3, p. 103]). This paper deals with the automatic extraction of tempo, and presents an elegant tempo/pace measure based on two relatively simple computable features; shot length and motion from digital movies and videos.

2. PREVIOUS WORK

While film grammar is mentioned in [5] with reference to the signatures some different types of dramatic events (e.g., release of tension) will leave, [5] is restricted to merely differentiating between the 3 chosen scene types. What our work seeks to do is, in a sense, to bridge the gap between those approaches [6, 7, 8, 9] that only perform low-level processing of videos and those [5, 10, 11] that rely heavily on extensive domain modeling to discriminate limited predefined events. Movie tempo, derived in this paper, can be seen to be both high-level and fundamental (therefore widely applicable), yet manifest in such a way as to be computationally tractable. This work thus moves away from sin-

gle frame/shot study and emphasizes across shot analysis for extraction of meanings that link shots.

3. NOVEL APPROACH TO MOVIE TEMPO EXTRACTION

Tempo or *pace* is a term that is broadly and often interchangeably used in video understanding and therefore in this paper as well. Pace often refers to perceived speed while tempo refers to the perceived duration and since our work derives both, we will use them together. Tempo/pace carries with it the important notions of time and speed and its definition reflects the complexity of the domain to which it is applied.

3.1. Establishing Tempo

How is tempo made manifest in film? How does a director create the desired pace? One way is by using the cinematic technique of *montage*. Montage, also known as editing, is “a dialectical process that creates a third meaning out of the adjacent shots” and has the ability to “bend the time line of a film” [2, p. 183,185]. Essentially, the director controls the speed at which a viewer’s attention is directed and thus impacts on her appreciation of the tempo of a piece of video.

A second way that tempo is manifest in film is through the level of motion or dynamics. Both camera motion and object motion impact on a viewer’s estimation of the pace of a piece of video. This is because motion can influence the viewer’s attention with more or less haste and strength.

There are many other elements which feed into this concept of tempo, music being another major contributor (along with the story). We will limit our consideration of tempo and pace to the factors of montage and motion in this paper for the following reasons: (i) The characteristic features of both montage and motion lend themselves well to automatic computation. (ii) They, together, form the major contribution to pace [3, p. 103].

Our approach is predicated on the notion that film sections of differing pace will leave distinct marks on the attributes of shot length and motion, and hence may be detected and classified based on those fundamental primitives. First we extract motion and shot lengths in a film. We then combine them in a novel fashion, to yield a continuous measure of pace for the course of the film. The input to our analysis is a compressed digital movie or TV program in MPEG-1 format.

3.2. The Tempo/Pace Function

We extract camera pan and tilt motion between successive frames in the input video stream using software implementing the qualitative motion estimation algorithm of [12]. The raw pan and tilt computed are then filtered of anomalous

values and smoothed with a sliding window. We also generate an index of shot boundaries (specifically *cuts*) by means of the commercial software *WebFlix*. The generated shot index is output as a series of start and stop frames.

While a simple classification scheme, e.g., a decision tree, based on these features would correctly categorize the really fast and really slow sections in a movie, it breaks down with sections that are neither decidedly fast nor slow [13]. The difficulty with this approach is first the issue of resolution. If the sections are too short we risk anomalous results; too long and we risk smoothing over subsections of markedly different tempo. Secondly, an objective decision about a section’s absolute pace is difficult. It is much easier to say “faster” than “fast” for example, and decisions can be affected by non-pace factors such as the emotional content of the section under consideration which relate more to higher semantic constructs such as tone or mood.

A more reliable tempo indicator would address the resolution issue and offer a more intuitive feel for the pace of a section within the context of the given film (i.e., would offer more relational information than a simple binary ordinal classification). Our solution has both of these desired attributes.

The continuous tempo/pace function, $\mathbf{P}(\mathbf{n})$ is defined as:

$$\mathbf{P}(\mathbf{n}) = \frac{\alpha(\text{med}_s - s(\mathbf{n}))}{\sigma_s} + \frac{\beta(m(\mathbf{n}) - \mu_m)}{\sigma_m}, \quad (1)$$

where s refers to shot length in frames, m to motion magnitude, and \mathbf{n} to shot number. The average motion magnitude is computed first for each shot, where the motion magnitude is simply the absolute value of the sum of the pan and tilt values for a given frame pair in the shot. Shot length, in frames (assuming a 25 frame/s rate), is also calculated for each shot. In addition to the per shot data, the mean, μ and the standard deviation, σ of these features are calculated for the entire film, along with the overall shot median, med_s . The weights α and β , are given values of 1, effectively assuming that both shot length and motion contribute equally to the perception of pace for a given film. Study of other weighting schemes is underway.

We also smooth $\mathbf{P}(\mathbf{n})$ with a Gaussian filter for two reasons. First it reflects our knowledge that directors generally do not make drastic pace changes in single or small numbers of shots. Secondly it also helps, in a very simple fashion, mimic the process of human perception of pace in that pace has a certain inertia to it due to memory retention of preceeding shots. That is, pace is a function of a neighbourhood of shots. The amount of smoothing changes the resolution of the tempo indication and correspondingly, the level at which pace features may be extracted.

Figure 1 is a plot of $\mathbf{P}(\mathbf{n})$ derived for the movie, *The Lost World Jurassic Park*. The zero axis in this plot may be roughly considered as the average pace mark for the film.

The plot offers a rough feel for the tempo profile of this adventure movie. It is apparent that there are about six alternating story sections of vastly differing pace. These correspond to a shortening pattern of conflict with the dinosaurs, and bridging breathing space crafted by the director.

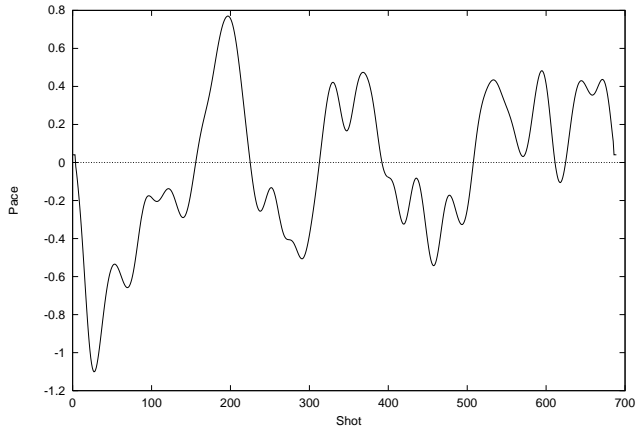


Fig. 1. Tempo plot for The Lost World.

3.3. Detection of Dramatic Story Sections and Events

Using this continuous measure of tempo, we locate the edges of the function $P(n)$ as they are good indicators of important pace transitions. Significant pace changes often occur across the boundary of story elements, and are often precipitated by events of high dramatic import in the story. A zero-crossing based edge detection process is carried out at multiple resolutions to extract locations of these changes.

Edges from the zero-crossings of the pace function are located using Deriche’s recursive filtering algorithm [14]. This multi-scale edge detection algorithm is parameterized by Σ , which determines the slope of the target edges. Larger Σ detects edges of smaller slope (more gradual) and vice versa. A threshold (τ) is applied to the resultant output of the algorithm to filter edges; the higher the threshold the fewer and larger the edges detected, and vice versa. The parameters used for the edge detection process are as follows: (i) $\Sigma = 30$, high $\tau (\pm 1.7\sigma)$ (of edge output): to locate significant, gradual pace transitions, (ii) $\Sigma = 30$, low $\tau (\pm 1\sigma)$: to locate all gradual pace transitions (large and small), (iii) $\Sigma = 0.5$, high $\tau (\pm 2\sigma)$: to locate significant, sharp pace transitions, and (iv) $\Sigma = 0.5$, low $\tau (\pm 0.8\sigma)$: to locate all sharp pace transitions (large and small).

Thus, four rounds of edge detection were applied to each film examined. *Large pace transitions* are targeted with a high threshold, and the resulting edges are designated “*story sections*” (roughly equivalent to “*scenes*”, although the definition of a scene is somewhat elusive in the context of film (see [2, p. 130] for a discussion of a “*scene*” in a modern film)). This label, albeit somewhat arbitrary, is useful

in terms of presenting the results of the edge detection process. *Small pace transitions* are accordingly called “*events*” since they are generally associated with localized events as opposed to changes of the order of story element size.

4. EXPERIMENTAL RESULTS

Results from one of the four movies analyzed, Titanic, Lethal Weapon 2, Colour Purple, and The Lost World Jurassic Park is presented here.

The sequel to Jurassic Park, The Lost World is full of more of the same as the InGen company seeks to capture more dinosaurs from the island known as “site B”. They end up struggling for life and in the end bring havoc to the mainland in the form of the mighty T-Rex. Figure 2 contains the pace plot of a section of the movie along with the edges determined, and Table 1 matches the edges with the story line.

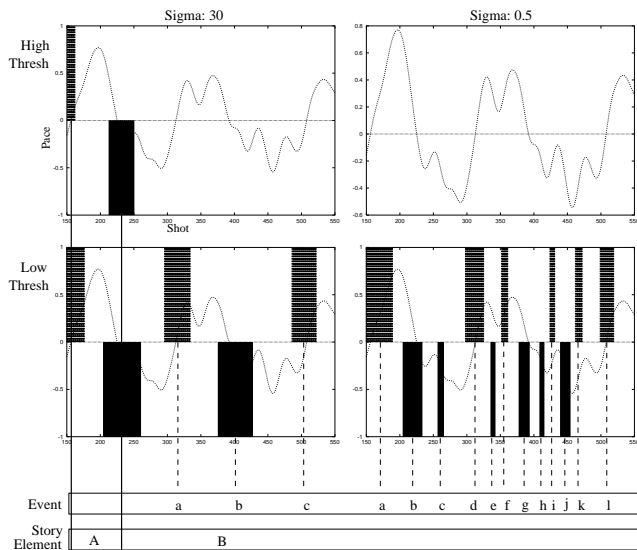


Fig. 2. Results of edge detection on pace flow and corresponding story sections and events from The Lost World.

In Table 1, gradual edge A, labelled “InGen dino hunt”, occurs at the point in the film where the InGen company employees reach the island to begin their dinosaur capture campaign. This signals a marked ramp up in the tempo of the movie as the exploits of these game hunters are shown, and is faithfully captured by the tempo measure. An example of a smaller edge indicating a dramatic local event, consider edge e, labelled as “Girl saved”. This small negative edge coincides with the partial rescue of the heroine from the truck dangling over the edge of a cliff. In the context of the T-Rex scene it serves to offer a momentary breather before the onset of yet another attack, but is duly signalled by our function and extracted as an event of interest.

In our experiments, this movie suffered the most from

Gradual Edge, Σ : 30		Sharp Edge, Σ : 0.5
<i>Story Element</i>		
A	InGen dino hunt	
B	Calm before T-Rex	
<i>Event</i>		
a	T-Rexs get angry	InGen hunters arrive
b	Rescue from dangling truck	finish the hunt
c	Attacked by raptors	Dinos are set free
d		T-Rexs get angry
e		Girl saved
f		Truck starts to slide again
g		Rescue from dangling truck
h		The group discusses
i		Man leaves group
j		...and gets attacked by dinos
k		T-Rex shows up again
l		Attacked by raptors

Table 1. Labelled story sections and events identified from tempo changes in *The Lost World*.

the misses (false negatives) during the shot boundary detection phase since a large portion of the movie is filmed at night. Webflix has been found to report both false negatives and false positives under conditions relating to dark sections of film (eg. some night scenes) and to periods of intense motion respectively. Within the context of the movie it does not effect the pace extraction to a great extent, but cross movie comparisons (particularly overall shot statistics) do suffer.

Three other movies, *Titanic*, *Lethal Weapon 2* and *Colour Purple*, were also subjected to the same analysis. Results were measured against a ground truth established from careful manual analysis of each movie. Overall the computation of $P(n)$ and subsequent edge detection succeeded in discovering nearly all actual distinct tempo changes. The list of located edges in the movies examined [13] serves as a useful and reliable index into the dramatic development and narration of the story.

A careful analysis of the performance shows that factors leading to incorrect edge detection include but are not limited to, poor shot indices, motion problems relating to scene depth, and poorly lit sections. In some cases estimated and reported edges differ for no other reason than that the resolution of the edge extraction process was too low (i.e., a function of the experiment, not the pace measure).

5. CONCLUSIONS

Taking advantage of film grammar, this work has sought to characterize the notion of movie tempo and pace as ex-

pressed by the indicators of shot length and motion, to produce a novel continuous measure of tempo. Our experimental results have demonstrated that the expressive element of tempo can be extracted, and the tempo function, together with its edges offer a rough feel for the pace changes in a movie from a quick glance. Further to this, it has also been shown that tempo is an effective attribute to compute as it offers pointers to higher level semantic constructs such as dramatic events and important story elements.

6. REFERENCES

- [1] D. Arijon, *Grammar of the film language*, Silman-James Press, 1976.
- [2] J. Monaco, *How to read a film: The Art, Technology, Language, History and Theory of Film and Media*, Oxford University Press, 1981.
- [3] T. Sobchack and V. Sobchack, *An introduction to film*, Scot, Foresman and Company, 1987.
- [4] Encyclopedia Britannica, "Encyclopedia Britannica Online," 1999.
- [5] A. Yoshitaka, T. Ishii, M. Hirakawa, and T. Ichikawa, "Content-based retrieval of video data by the grammar of film," in *IEEE Symposium on Visual Languages, Capri, Italy*, 1997.
- [6] R. Lienhart, W. Effelsberg, and R. Jain, "Visualgrep: A systematic method to compare and retrieve video sequences," Tech. Rep., The University of Mannheim, 1997.
- [7] R. Hammoud, L. Chen, and D. Fontaine, "An extensible spatial-temporal model for semantic video segmentation," in *1st Int. Forum on Multimedia and Image Processing, IFMCP '98, Anchorage, Alaska*, 1998.
- [8] W. Mahdi, L. Chen, and D. Fontaine, "Improving the spatial-temporal clue based segmentation by the use of rhythm," in *Second European Conference, ECDL '98*, 1998.
- [9] A.D. Doulami, Y.S. Avrithis, N.D. Doulami, and S.D. Kollias, "Interactive content-based retrieval in video databases using fuzzy classification and relevance feedback," in *IEEE Int. Conf. Multimedia Computing and Systems*, 1999.
- [10] N. Vasconcelos and A. Lippman, "Bayesian modeling of video editing and structure: Semantic features for video summarization and browsing," in *ICIP'98, Chicago*, 1998.
- [11] Stephen S. Intille and Aaron F. Bobick, "Closed-world tracking," in *Proceedings of the 5th International Conference on Computer Vision*, Cambridge, MA, June 1995, pp. 672–678.
- [12] M.V. Srinivasan, S. Venkatesh, and R. Hosie, "Qualitative extraction of camera parameters," *Pattern Recognition*, vol. 30, no. 4, 1997.
- [13] B. Adams, C. Dorai, and S. Venkatesh, "Towards automatic extraction of expressive elements from motion pictures: Tempo," in *IEEE International Conference on Multimedia and Expo, To appear*, New York City, NY, USA, July 2000.
- [14] R. Deriche, "Recursively implementing the Gaussian and its derivatives," in *ICIP'92, Proc. 2nd Singapore Int. Conf. on Image Processing*, 1992, pp. 263–267.