Deakin Research Online

This is the published version:

Adams, Brett, Dorai, Chitra and Venkatesh, Svetha 2002, Toward automatic extraction of expressive elements from motion pictures : tempo, *IEEE transactions on multimedia*, vol. 4, no. 4, pp. 472-481.

Available from Deakin Research Online:

http://hdl.handle.net/10536/DRO/DU:30044282

Reproduced with the kind permissions of the copyright owner.

Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Copyright : 2002, IEEE

Toward Automatic Extraction of Expressive Elements From Motion Pictures: Tempo

Brett Adams, Chitra Dorai, *Senior Member, IEEE*, and Svetha Venkatesh, *Senior Member, IEEE*

***Abstract*—This paper addresses the challenge of bridging the *semantic gap* that exists between the simplicity of features that can be currently computed in automated content indexing systems and the richness of semantics in user queries posed for media search and retrieval. It proposes a unique computational approach to extraction of expressive elements of motion pictures for deriving high-level semantics of stories portrayed, thus enabling rich video annotation and interpretation. This approach, motivated and directed by the existing cinematic conventions known as *film grammar*, as a first step toward demonstrating its effectiveness, uses the attributes of motion and shot length to define and compute a novel measure of *tempo* of a movie. Tempo flow plots are defined and derived for a number of full-length movies and edge analysis is performed leading to the extraction of dramatic story sections and events signaled by their unique tempo. The results confirm tempo as a useful high-level semantic construct in its own right and a promising component of others such as rhythm, tone or mood of a film. In addition to the development of this computable tempo measure, a study is conducted as to the usefulness of biasing it toward either of its constituents, namely, motion or shot length. Finally, a refinement is made to the shot length normalizing mechanism, driven by the peculiar characteristics of shot length distribution exhibited by movies. Results of these additional studies, and possible applications and limitations are discussed.**

***Index Terms*—Content-based search and annotation, dramatic sections, events, expressive elements, film grammar, film pace, genres, mood, motion pictures, semantic gap, styles, subjective time, tempo, video archive.**

I. Introduction

INCREASED use of the rich video medium has highlighted the inadequacy of current tools for automated content understanding. Such tools are vital for effective indexing and browsing of any body of video data. While many research attempts have sought solutions to the problem by extending the interrogative techniques of the relatively mature textual/image query, it has become apparent that these fall short in mining meaning from the unique modes open to the video medium. The sharp discontinuity, referred to as the *semantic gap* [1], between the simplicity of features that can be currently computed in automated content management systems, both commercial and research prototypes, and the richness of user queries encountered in media search and navigation makes user acceptance and adoption of these systems very difficult.

What is needed is a recognition that the problem requires a paradigm shift. Such was the case for those who recognized that simple text annotation was not adequate to the needs of image query. This realization resulted in such systems as query by image content (QBIC) [2] and photobook [3]. In essence, this was a realization that a query language must align, as best can be done, human perception of a medium and computer perception of the same. For images this has seen texture, color, and shape (so far) become the currency of use. What are the analogous elements of video?

It is here that the existence of a form of film "grammar" would seem to offer some hope [4], [5, p. 119] [6, p. 189]. Far from being the strict rules of written language communication, *film grammar*, which embodies video production knowledge, is found more in history of use and is descriptive rather than prescriptive. Potentially it elucidates on the apparently obscure relationships that exist between the many cinematic techniques employed by a director and their semantic interpretation by a viewer.

We propose a unique approach, guided systematically by the grammar, to computationally determine the expressive elements of motion pictures conveyed by the manipulation of editing, lighting, camera movements, color, etc., for high-level video understanding and appreciation. Our work focuses on the extraction of high-level semantics associated with the *expressive elements* and the *form* of story narration in films. It differs from many recent approaches in that while others have sought to model very specific events occurring in a specific domain, our research attempts to understand the "expressiveness" of the medium and the thematic units (high-paced section, tranquil event, etc.) highlighted by the expressions, that are pervasive regardless of the domain of the story.

One expressive element that forms part of this *film grammar* and movie understanding is that of *tempo* or *pace* influencing the sense of a story's *experienced* time. Sobchack says that "[Pace] is usually created chiefly by the rhythm of editing and by the pace of motion within the frame" [6, p. 103]. This paper is concerned with the automatic extraction of tempo/pace from video and its implications to high-level semantic analysis.

The remainder of this paper is structured in the following manner. Section II is concerned with the nature of *film grammar* and its importance as the guiding basis of any attempt at automated film understanding. Section III will examine past work in the area of semantic analysis of film and video from the perspective of the degree to which guidance and inferences have been drawn from *film grammar*. Section IV forms the major contribution of this paper, detailing the derivation of a film tempo/pace measure, and the utilization of this measure

Manuscript received March 8, 2002. The associate editor coordinating the review of this paper and approving it for publication was Dr. Ahmed Tewfik.

B. Adams and S. Venkatesh are with the Department of Computer Science, Curtin University of Technology, Perth, WA, Australia (e-mail: adamsb@cs.curtin.edu.au; svetha@cs.curtin.edu.au).

C. Dorai is with the IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: dorai@watson.ibm.com).

Digital Object Identifier 10.1109/TMM.2002.802016

1520-9210/02$17.00 © 2002 IEEE

for the extraction of dramatic sections and events from film. Section V contains novel improvements and insights to the tempo function derived in Section IV in the form of biasing toward different styles/genres (Section V-A), and refinement resulting from greater faithfulness to the factors that influence shot length distribution (Section V-B). A discussion follows in Section VI with a conclusion to the paper found in Section VII. Some of the material in this paper has been presented at conferences [7]–[9].

II. Film Grammar

The first step in any process of understanding "data" should be a survey of the forces that go toward shaping that data. If a written text is known to have been produced by a million monkeys on a million typewriters, then this text has implications far different than if the text were known to have originated with a human author (e.g., the Bard himself). In a sense, defining the set of forces that are considered to be candidates for the manifestation of given data, is in fact a definition of the domain at hand.

Having defined our domain as Film (excluding perhaps the most extreme types of art house film, though even these have a grammar all their own), we inherit a set of candidate "forces" that are collectively termed *film grammar*. *Film grammar* is defined in [4, p. 2] as being "comprised of a body of 'rules' and conventions" that "are a product of experimentation, an accumulation of solutions found by everyday practice of the craft," and results from the fact that films are crafted, built, shaped to convey a purpose (whatever that may be). For example, it has been found that a hand-held camera is a powerful expressive tool for evoking a subjective feel from a piece of film. A better known example is the use of a low camera angle to invoke fear.

If one takes the analogy of film as a meaningful text, similar to a written text, just as understanding say English grammar is essential to understanding an English text, an understanding of *film grammar* is essential to understanding film (even at its most fundamental level).

A note of caution should be sounded at this point. *Film grammar* is very much descriptive rather than prescriptive in nature. As such it should be used with a degree of caution and flexibility. There is a strong recognition, however, that filmmakers generally ignore the tenets of *film grammar* at the risk of confusing their audience, and thus reduce the ability of their film to communicate their desired intent (unless of course, the filmmaker is seeking to communicate confusion).

III. Background

A survey of work in the field of automated film understanding will now be provided from the viewpoint of the degree to which *film grammar* has been utilized by the outlined approaches.

Lienhart *et al.* use a feature set that includes "motion intensity," color atmosphere, lightness, orientation, frontal faces, and type of framing, for comparison of film sequences from a hierarchy of different temporal resolutions [10]. They quote [11] on the definition of a scene, and attempt to link some of their features with the film literature (e.g., type of framing).

Fischer *et al.* developed a genre recognition system based upon a feature set similarly comprised [12]. A best match for the formed "style attributes" was then made to determine which genre the video section was from. The system discriminated between five different genres, namely, Newscast, Car Racing, Tennis, Commercials, and Animated Cartoons. The point is made that "the amount of camera motion and object motion in a film is an important style attribute;" however, the style profiles are largely formed from an inductive consideration of the data rather than from systematic inferences drawn from *film/TV grammar.*

Clustering shots into semantically related groups based on a number of visual primitives including optic flow was undertaken by Hammoud *et al.* [13] (with the basis for shot similarity being implicitly assumed). Clusters are then further refined based on temporal relations of member shots. This system was further extended in [14] to overcome the problem of single shot scenes via the assumption that shot length is normally distributed. However, a consideration of causes and real movie data indicate that a tailed distribution such as Weibull or Lognormal would be appropriate (see Section V-B, or [15, p. 224], [16]).

Vasconcelos *et al.* [17] employs a Bayesian network fed by three visual sensors, namely, motion energy, skin, and texture energy. Each sensor provides an input to the Bayesian network to help discriminate a given shot in terms of four categories of content: action, close-up, crowd, and setting. The authors do state that they are seeking to exploit production codes, and provide the example of the use of close-ups in romantic movies, but do not provide a strong theoretical link between the choice of sensors and the inferred content. In earlier work, the same authors recognized that different genres leave their signatures on overall shot and dynamicity [16]. They plot each film on a continuous scale of overall activity versus shot length and separate films into romance/comedy and action. "Activity" is calculated here using a "tangent distance" as an attempt to discriminate this measure from unwanted lighting change or camera motion.

Yoshitaka *et al.* use a mixture of shot length, luminance change (shot dynamics) and image similarity to detect from among three types of scene, namely, conversation, increasing tension, and hard action [18]. While *film grammar* is mentioned with reference to the signatures some different types of dramatic events (e.g., release of tension) will leave, their work is restricted to differentiating between the three chosen scene types.

Doulamis employs shot detection, key frame/shot selection, object segmentation and tracking, and fuzzy classification coupled with user feedback to perform refined queries [19].

Sundaram and Chang [20], [21] extend and modify an idea first proposed by [22] for scene segmentation. Here the basis for visual shot similarity is whether or not a group of shots is "chromatically consistent" (i.e., color), and audio features are incorporated for the segmentation of "audio scenes." The data from the visual and aural modes are then analyzed within the context of a memory/attention span model [22] for the purpose of finding likely segmentations, or singleton events.

It is worth noting the attempt by Radev *et al.* [23] at providing a general film model. While not offering any methods for automating a classification system in accordance with their model, they do show in concept the possibility of deriving such models from the film literature.

All of the solutions detailed above tend toward falling into one of two (somewhat overlapping) categories.

1) The first category is generally concerned with finding scene or sequence *boundaries* or *similar sections* of film. These approaches seek to extract as much information as possible from the video source. All further processing is then founded on the basis of image/audio similarity measures. Shots that are similar in terms of the multitude of extracted features are considered to be semantically similar. Further investigation is then optionally carried out on the discovered "semantic units" in terms of some *a priori* temporal model(s). The term *scene* here is generally conceived of in a limited fashion.[1] In many ways the scene/sequence construction is little codified, and often the elements that bind a series of shots into a meaningful "sequence" are not color, motion, etc.

   The essential problem with most examples of this approach is how a fundamental question is answered, namely: What constitutes shot similarity? To answer this, we must pose it as "What shots does the viewer find similar?" which is largely answered by determining what shots the filmmaker intends to be semantically similar. That is, it is the filmmaker who crafts a series of shots into a unified, meaningful whole. *Film grammar* tells us about the variety of ways that this may be achieved—music, a character, a theme, and how filmmakers reinforce this unity. A better appreciation of *film grammar* will, therefore, push back the limits of tools so far founded upon a simplistic view of scene structure.

2) The second approach is generally concerned with determining the *nature* of all or part of a given film's *content* and involves discriminating between predefined categories of shot/sequence/film based upon a careful selection of low-level features that map well to high-level features for the given categorization problem. The focus here is on spotting "useful" features as opposed to trying to completely reassemble the full spatio-temporal nature of the video contents.

   The problem with many of the approaches in this category is that the domain must be limited to an impractical degree or else the carefully selected feature mappings break down. As a consequence, the query/annotation scope is limited to that set of high-level interpretations allowed. An understanding of *film grammar* is essential here also, as it details fundamental features that are ubiquitous, yet are strongly linked to high-level interpretations.

What our work seeks to do is, in a sense, somewhere between the above two approaches. On one hand it is recognized that the problem of reconstructing the full spatio-temporal source of video is a very difficult and computationally expensive problem. On the other hand, selective feature extraction results in a query scope that is very limited in proportion to the high-level indications sought. Fundamentally our work is a call for the *systematic* application of *film grammar* to the building of tools for the purpose of extracting semantic constructs from film. As an example of the process we will consider one aspect in particular. The expressive element, tempo or pace, as discussed in our paper can be seen to be both fundamental (therefore widely applicable), yet manifest in such a way as to be computationally inexpensive; in effect answering the above two challenges.

IV. Computational Approach to Movie Tempo Extraction

This section will define tempo, and detail the process of drawing inferences about it from the film literature leading to the construction of a computable entity.

A. *Defining Tempo*

Tempo or pace is a term that is broadly and often interchangeably used in film studies and therefore in this paper as well.[2] A helpful definition in this context might be "rate of performance or delivery." Tempo carries with it the important notions of time and speed and its definition reflects the complexity of the domain to which it is applied. A runner has a simple velocity, music has a tempo and rhythm, a time signature that speaks to beat and bar. Video can be quite complex including both of the above at once.

How is tempo made manifest in film? or more precisely: How does a director manipulate time and speed in a film to create a desired tempo? 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" [5 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 video. This is because motion, like montage, 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 (currently under investigation as another component of tempo), and story. We will limit our consideration of tempo/pace to the factors of montage and motion in this paper for the following reasons.

1) The characteristic features of both montage and motion lend themselves well to automatic extraction.
2) Together they form the major contribution to pace [6, p. 103].

[1] When applied to modern filmmaking "the term *scene* is useful but not precise" [5, p. 130]. The term is borrowed from the French classical theatre and had a precise beginning and ending (arrival and departure of character). However, when applied to film, the term undergoes a widening of use in proportion to the opportunities available to the roving camera and other cinematic techniques. "Few shooting scripts are divided into 'scenes.' Scene can refer to a tableau (e.g., a sunset), a place, or an action; preferred terms are shot and sequence, though one still speaks of a 'love scene'" [24, p. 107–113].

[2] Zettl [25, p. 249] distinguishes between pace "the perceived speed of the overall event," and tempo, "the perceived duration of the individual event sections". As one may be derived from the other, and as Zettl labels the distinction as "confusing and bothersome," the terms will be used interchangeably in this paper.

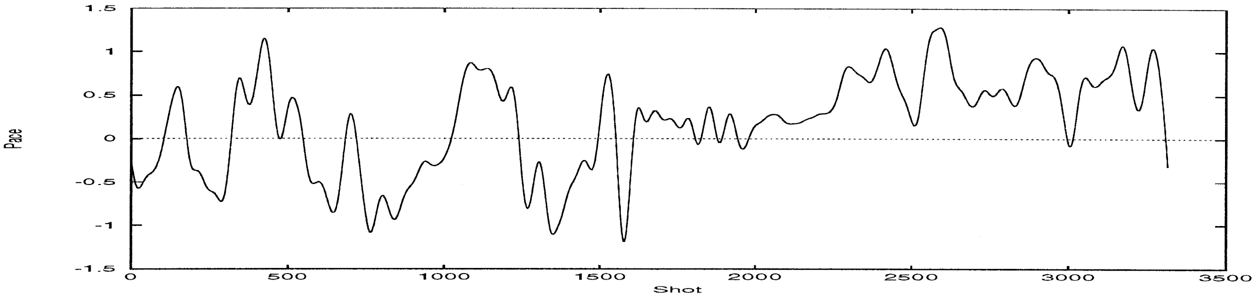

Fig. 1. Plot of the pace function for *Titanic*.

B. Extracting the Required Data

The input to our analysis is a compressed movie or TV program in MPEG-1 or other format. Extraction of camera pan and tilt between successive frames was performed on the input video stream with software implementing the qualitative motion estimation algorithm of [26]. The raw pan and tilt computed were then filtered of anomalous values and smoothed with a sliding averaging window of 25 frames.

An index of shot boundaries (specifically *cuts*) is created from the video by means of the commercial software *WebFlix* [27]. Although imperfect, it has been found overall, to do an adequate job of automatic shot detection. The generated shot index is output as a series of start and stop frames. Shots of length smaller than ten frames (under half a second) are merged, as they are deemed to be false positives.

C. A Tempo Formulation

As a preliminary validation of our approach founded on motion and shot length, we carried out a classification exercise [7] with data from three mainstream movies, namely, *Titanic*, *Lethal Weapon 2*, and *The Color Purple*. A decision tree classifier was built for automatic categorization of movie sections (of the order of 1000–5000 frames) as either *fast* or *slow* in tempo, using the C4.5 software [28] and using the motion and shot statistics as features. The classification results proved promising, in that they confirmed our hypothesis tying tempo to shot length and motion, but inadequate as a solution for faithfully representing tempo. What is desired is a *continuous measure* that captures the ebb and flow of tempo within a film, that both addresses the resolution issue and allows a more intuitive feel for the relative tempo of a section within the context of the given film (i.e., would offer more relational information than a simple binary ordinal classification).

We formulate a new pace/tempo function based on motion and shot length, exhibiting the desired characteristics. The average motion magnitude is computed for each film shot, where the motion magnitude is simply the absolute value of the sum of the pan and tilt values for a given frame pair. Shot length, in frames (assuming a 25 frame/s rate), is also calculated for each shot.

Pace is initially defined as

$$P(n) = \frac{\alpha(\mathrm{med}_s - s(n))}{\sigma_s} + \frac{\beta(m(n) - \mu_m)}{\sigma_m} \qquad (1)$$

where

s	shot length in frames;
m	motion magnitude;
n	shot number;
σ_s and σ_m	standard deviation of shot length and motion, respectively;
μ_m and med_s	motion and median of shot length, respectively.

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. Other weighting as well as feature normalization schemes are investigated in Section V.

The pace function $P(n)$ is then smoothed with a Gaussian filter with a size 9 window ($\sigma = 1.5$). Besides smoothing the data this process is desired for two reasons. First it reflects our knowledge that directors generally do not continue to make drastic pace changes in single or small numbers of shots, unless motivated by rare narrative requirements. Second, 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 preceding shots. That is, pace is a function of a neighborhood of shots. As anticipated, the amount of smoothing changes the resolution of the tempo indication and correspondingly, the level at which pace flow features may be extracted.

Fig. 1 is a plot of $P(n)$ for *Titanic*, with the Gaussian smoothing applied 100 times. The zero axis in this plot may be roughly considered as the average pace mark for the film. The first half of the plot encompasses the day before Titanic sinks, up to the point where the iceberg strikes. The second half of the plot depicts from that time until the ship sinks, and is conspicuous by the marked increase in pace (staying above the reference pace mostly) accompanying it.

D. Using $P(n)$—An Event and Dramatic Section Boundary Detector

Given this continuous measure of tempo there are many features that one might extract that would be useful. We have initially chosen to locate the edges of the function $P(n)$ as it is a relatively straight forward task, and more importantly is a good indicator of important events. Significant pace changes often occur across the boundary of story elements, and are often precipitated by events of high dramatic import in the story. Edge analysis is performed to determine locations of these changes.

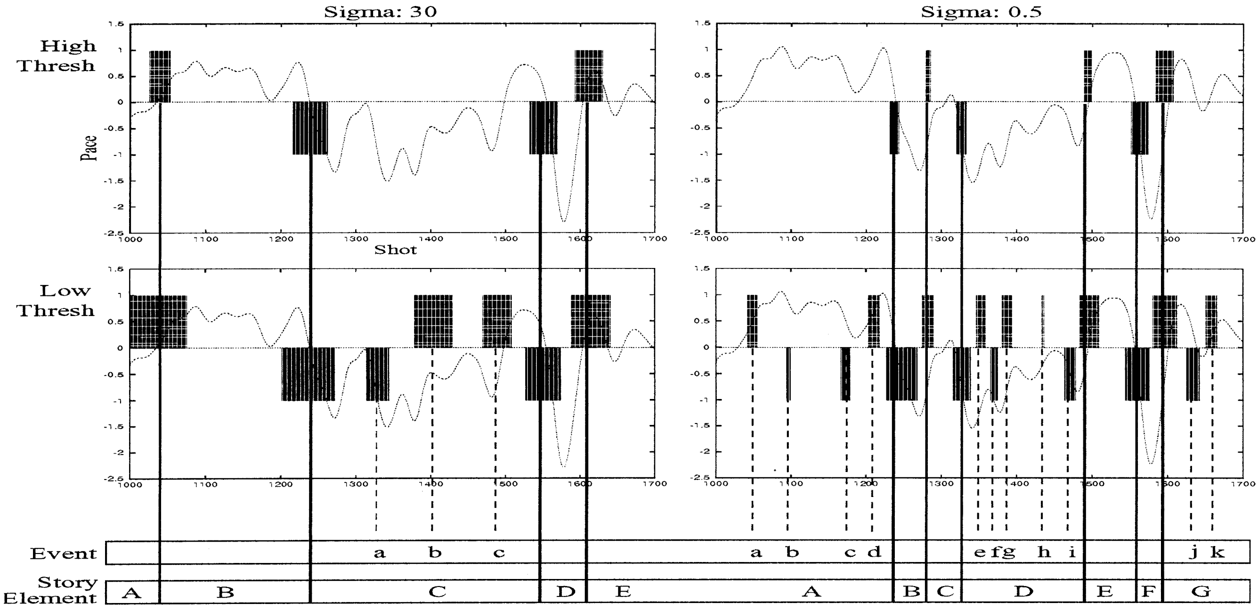

Fig. 2. Results of edge detection on pace flow and corresponding story sections and events from *Titanic*.

Edges of the pace function are detected using Deriche's recursive filtering algorithm [29]. This multiscale edge detection algorithm is parameterized by Σ, which determines the slope of the target edges. Larger Σ detects edges of smaller slope (more gradual change) 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:

1) $\Sigma = 30$, high τ ($\pm 1.7\sigma$ of edge output): to locate significant, gradual pace transitions;
2) $\Sigma = 30$, low $\tau\,(\pm 1\sigma)$: to locate all gradual pace transitions (large and small);
3) $\Sigma = 0.5$, high $\tau\,(\pm 2\sigma)$: to locate significant, sharp pace transitions;
4) $\Sigma = 0.5$, 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 to bound *story sections*. This label is somewhat arbitrary as large transitions do not always indicate a change of story element and vice versa; however, it is useful in terms of presenting the results of the edge detection process. *Small pace transitions* are accordingly called *events* due to the fact that such transitions are generally associated with localized events as opposed to changes of the order of story element size.

1) Experimental Results: First, results from *Titanic*, one of a number of movies analyzed are presented in detail for the purpose of demonstration. *Titanic* is a love story centered around the event of the sinking of the Titanic. Fig. 2 shows the pace plot of a section of the movie (from the third class party up to the point where the iceberg hits) with located edges indicated for each of the four Σ/τ combinations used, and Table I matches each automatically discovered edge to a brief description of the story section bounded by, or the dramatic *event* coinciding with the discovered edges.

TABLE I
LABELLED STORY SECTIONS AND EVENTS IDENTIFIED FROM TEMPO CHANGES IN TITANIC (SEE Fig. 2)

	Gradual Edge, Σ: 30	**Sharp Edge, Σ: 0.5**
Story Element detected (high thresh)		
A	1st class dinner	Day before sinking
B	3rd class party	Rose in trouble
C	The next day	Rose has to decide
D	Calm before the storm	Rose chooses Jack
E	Titanic is sinking	Jack and Rose chased
F		Calm before the storm
G		Titanic is sinking
Event detected (low thresh)		
a	Jack and Rose D&M	Dancing begins
b	Jack sketches Rose	Jack partners with Rose
c	Jack and Rose chased!	Rose stands on toes
d		Dancing again
e		Rose chooses Jack
f		At the bow of Titanic
g		Present day to Rose and Jack back at room
h		Jack sketches Rose
i		From sketch to pres. day
j		Iceberg seen, tense wait
k		Iceberg hits

Consider Table I. The first large gradual edge reported occurs at the transition between the story elements A and B, labeled as "first class dinner" and "third class party," respectively. The difference between the lives of the first class and third class people is a dominant theme throughout the movie and is expressed here by the stiffly formal nature of the former contrasted with the exuberance of the latter. As such it marks a large change in the pace which is duly signaled by our algorithm. The next large gradual edge occurs at the transition to the next story element C, labeled "the next day." This is a negative edge and marks the change of tempo that occurs as Rose is seen back in her first class life.

On a finer scale of sharp edges, the sharp negative edge j "tense wait after iceberg seen" occurs as the initial flurry at the

TABLE II
RESULTS OF TEMPO-BASED EDGE DETECTION IN MOVIES

Movie	Edges Found	False Neg.	False Pos.
Titanic	68	5	7
Lethal Weapon 2	17	4	1
Lost World	19	3	0
Color Purple	18	4	0

sight of the iceberg dies and the crew wait to see whether the Titanic will clear it or not. The next sharp positive edge k, "iceberg hits" coincides with the actual impact of the iceberg and the ramp up in tempo as the resulting damage is graphically portrayed in Fig. 2

2) Overall Results: Overall the computation of $P(n)$ and subsequent edge detection succeeded in discovering nearly all actual distinct tempo changes with good precision and recall. It should be noted that the edge experiment parameters were set once for all movies and underwent no subsequent tweaking. As such the experimental results represent an under estimation of the abilities of the measure. The resulting list of located edges in all four movies [*Titanic (TT), Lethal Weapon 2 (LW2), The Lost World Jurassic Park (JP2),* and *The Color Purple (CP)*], shown in Table II serves as a useful and reliable index into the dramatic development and narration of the story.

Experience with this measure in the context of an interactive workbench software further confirms the formal results presented in Table II. $P(n)$ has demonstrated the ability to reflect the ebb and flow of film tempo and consequent dramatic indicators consistently across over 20 full-length movies to date.

V. Improvements and Insights

A. Flavors of Tempo

While it has been recognized that both motion and shot length contribute to the perception of pace, it is not as simple to make statements about their relative impact. *Film grammar* tells us that different directors and different genres will make use of these complementary techniques with differing emphasis.

Research using $P(n)$ has thus far used unit weights for α and β. This assumes that shot length and motion contribute equally to the perception of time. It is possible, however, that under certain circumstances, one or the other of these two impact more heavily on the audience perception of time. Such circumstances might include sparing use of a technique. For example, a director who makes minimal use of quick cutting techniques might do so only at particularly important junctures in a film's development. Conversely, a director might rely heavily upon one technique to clarify the story. An example of this might be films crafted by montage directors, who make use of shot characteristics (e.g., length and placement) for the purpose of carrying artistic expression to the exclusion of other cinematic factors like motion. This results in one technique being used to influence semantic interpretation, thus relegating other techniques such as motion to a position of secondary importance.

In order to explore this further, we carried out two experiments on the movies, *Color Purple* and *Lethal Weapon 2*. These movies were chosen as good representatives of a slow thoughtful movie and an action movie, respectively. Due to space restrictions we present only a brief analysis of the results, and refer the interested reader to [8].

The first experiment consisted of increasing the sensitivity of $P(n)$ to the *less used cinematic technique*. For LW2, this is shot length $(\alpha = 1.5, \beta = 0.5)$, and for CP this is motion $(\alpha = 0.5, \beta = 1.5)$. Edges were automatically detected on new pace measures. The results for LW2 saw an increase of edges to do with character/plot development and background rather than the dynamic action sequences that are the hallmark and raison d'etre of this genre of film. With unit weights many of these important edges are swallowed by the predominant, visually dynamic events, but their detection is desirable as it provides a useful index to the overarching story providing context for the action. Unlike LW2, the story told by CP is by nature "amotion." Its purpose is not to thrill the viewer with visually spectacular scenes, rather, the centerpiece of the story consists of the relationships that exist between the characters. Therefore, we have no clear trend being manifest with increasing shot motion's contribution, which fits with the fact that we are amplifying *semantic noise*.

The second experiment saw us increase the sensitivity of $P(n)$ to the *primary cinematic technique*. The weights α and β swapped their values accordingly. For LW2, this increased emphasis on motion's contribution resulted in a shift toward more action based events, those noted above as being the trademarks of this kind of film. We also found better resolution of different events within larger action sequences. The converse is true for CP. Here there was a swing toward dramatic events. Given the discussion above of the relatively minor role of motion in the fashioning of this story, it is to be expected that the result of desensitizing the pace function to motion is to more faithfully bring out the ebb and flow of the story. The results suggest possible application of tempo to genre, or event type, specific indexing, in either an automatic or interactive setting.

B. Enhancements to Shot Length Normalization

The pace function $P(n)$ is a composite of two shot features, namely, motion and shot length. Each feature is normalized such that a given value makes an intuitive contribution to the overall calculation of pace for a shot. The normalization process is undergirded by assumptions that should be noted here.

First, let us look at motion. With no impetus to proceed otherwise, it is assumed that motion is normally distributed and therefore the form of normalization seen in (1) is adopted for motion. The second component of the pace function is shot length which will be dealt with in detail in this section. Initially, it too is treated as being drawn from a normal distribution [see (1)]. However, an analysis of the raw data, and of the processes leading to the formation of shot length, suggest that a better model for normalizing shot length is required.

Vasconcelos *et al.* [30] makes the point that shot length appears to be adequately modeled by a member of the Weibull distribution family. Our experiments also confirm this, from the distributions of shot length data for movies from various genres such as TT, LW2, JP2, CP (and many others). These distributions are roughly Weibull. Fig. 3(a)–(c) shows several examples.

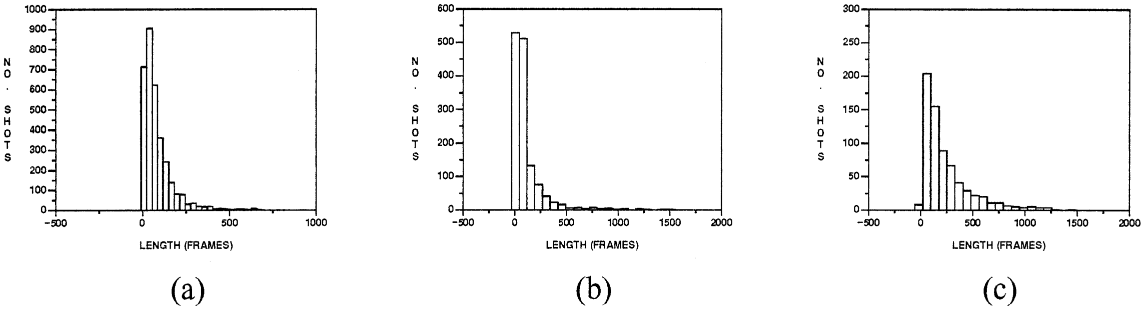

Fig. 3. Shot length distributions: (a) *Titanic*; (b) *Lethal Weapon 2*; and (c) *The Lost World Jurassic Park*.

This apparent distribution is also somewhat predictable from a consideration of human abilities and the movie making process as explained in the following subsection.

1) Shot Length Distribution Characteristics: For most situations there is a practical lower limit placed on shot length by the ability of a human viewer to adjust to and process the information of a new shot. It is a testimony to the art of directing that these transitions, for the most part, go unnoticed. Many of the tenets of *film grammar* pertain to exactly this criterion, e.g., the avoidance of crossing the eye-line, or motion and direction continuity between shots. The fact is that a movie is made up of a series of disjoint shots that would appear very confusing were it not for that fact that we are "trained" to interpret them [5, p. 121]. Reference [31] refers to studies that indicate that it can take between 0.5 s and 3 s for a viewer to adjust to a new shot, let alone absorb the contents of that shot. While this can sometimes be used as a creative tool ([31] notes the sequence in the film *Patriot Games* where a series of very short shots is used to create tension as Jack Ryan watches the terrorists killed on the other side of the world via satellite. It should be noted also, that there is very little new information to be assimilated in each of the consecutive shots), usually it is desired that the information in each shot be taken in and added to the growing story.

The upper limit to shot length is a lot more amorphous, whereas the lower limit has a large degree of inflexibility to it that is derived from its physical nature, the upper limit generally tends to be dependent on more subjective factors like audience interest levels and the degree of complexity of story to be captured. For example, many very long shots may lose audience interest, or be unable to aptly convey the desired information intended by the story teller.

The logistics of movie making also impacts on the possible shot makeup of a film. The more intricate the pattern of shots, the more the cost in time and money (although new editing and recording technology will affect this to a degree). Often a director has to resort to using significant pieces of the master (or cover) shot when planned interleaved shots are deemed unsatisfactory [5]. In other words, the final shot makeup of a film is not always, if ever, the creative ideal envisioned by the director/editor and hence will, to a degree, reflect these underlying processes in its manifest shot length distribution.

Given that shot length data exhibit something like a Weibull distribution, we address the following question: How do we then formulate the pace equation such that it makes comparable and intuitive contributions from the shot length to the output?

2) Appropriate Shot Length Normalization Model: In one sense the shape of the distribution for shot length does not tell us anything about the connection of a certain shot length to a certain perceived time. For this we need to consider perception of time goals (from the director's point of view), and reception (from the audience's point of view).

Zettl [25] offers some clues as to what a director is trying to achieve with different shot length ranges. Small shot lengths are manifest during the latter part of an increasing metric montage, or during a fast paced metric montage. The director's goal here is to intensify the event by an increasing, or high event density, achieved by means of rapid shot changes. The result being a fast paced section. Shots whose lengths lie closer to the overall median value are possibly the result of a great many factors. Rhythmic montage, medium paced metric montage, and many narrative restraints all result in shot lengths at or near the median. In one sense, this range of shots is the default range for maintaining audience interest levels. As such these shots assume the role of "midpoint" in a tempo estimation. Above the median, typically shot length can range up to a considerable maximum length. Shots of longer length have less of a role in decreasing event density as event clarification can be carried by other methods. That is, the role of shot length in relation to influencing audience perception of time becomes subordinated by the in-shot elements such as primary/secondary (i.e., object/camera) motion, story development, and dialogue etc. From an indication based purely on shot length these shots indicate a below normal tempo. Proportionally however, their influence should be considered to be decreasing based on the assumption that the in-shot parts may be being made the vehicle of tempo.

To summarize then, the role of shot length in affecting perceived time is most pronounced from the minimum shot length to somewhere above the median, but well short of the maximum shot length. Our shot normalization scheme should thus be most sensitive in these areas, and less so as shot length proceeds beyond this area. Added to this are some others factors that influence human perception of pace.

3) The Proposed Shot Length Normalization Scheme: We choose to separate the shot length data computed from a movie into two sections, shot lengths below and above the overall median.

The median is chosen as the "zero" point of the contribution of shot length to pace. Half of the shots have durations above this point, and half below (by definition). The median provides a more robust estimate of the average in the presence of outliers.

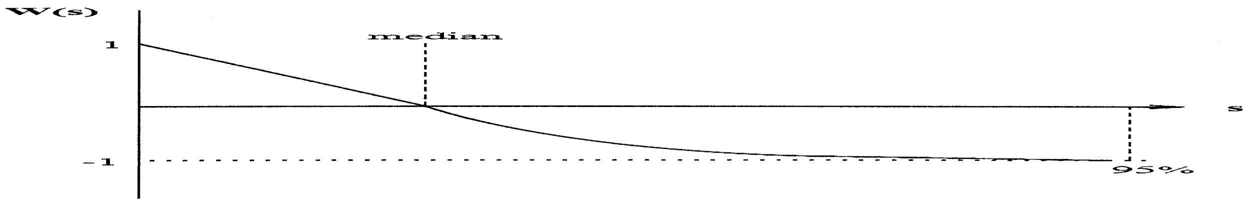

Fig. 4. New shot length normalization scheme for tempo computation.

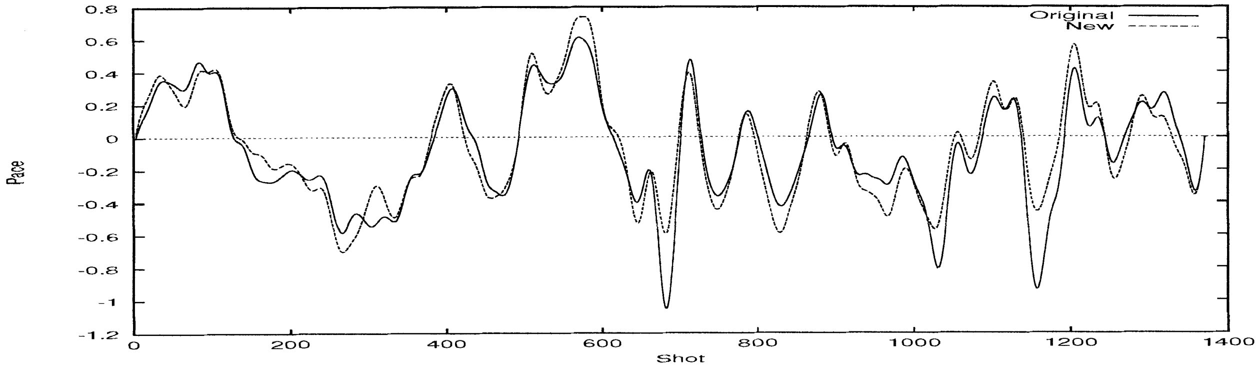

Fig. 5. Comparison of shot normalization schemes with the pace plots of LW2.

The sample space below the median is well contained (by the minimum shot length below, and the median above) and as such can be weighted with a simple linear model. The slope of the curve is chosen such that the minimum shot length coincides with a unit weighting (for symmetry).

Above the median is a different story. Shot length is theoretically unbounded, and in practice includes outliers far removed from the preponderance of data. A linear model would, in general, underweight the majority of data.

A better scheme would be to use the hazard function of an underlying Weibull model. Experiments with distribution fitting software [32] reports a beta of the order of 1.5, which results in a hazard function of the form $x^{1/2}$. Fitting this curve to reach -1 (for symmetry) at the 95% mark of shot lengths, for robustness, results in the following overall two part weighting scheme. This function has the property of being more sensitive near the median, but slows in gradient as shot length increases into the "longer" range. This is a desirable attribute given the above discussion of shot length ranges and their contribution to pace. Fig. 4 shows a plot of the new shot length normalization scheme $W(s(n))$, and the enhanced tempo function assumes the form

$$P(n) = \alpha(W(s(n))) + \frac{\beta(m(n) - \mu_m)}{\sigma_m}. \tag{2}$$

4) Results With the New $P(n)$: Fig. 5 shows what the new shot normalization scheme does to the pace and to the detected events of LW2. It shows that the new scheme fills in the large tempo drops that occur at shots 700 and 1200. These scenes ("boss talks to girl/Riggs, Murtaugh, and Leo in car," and "the calm before Riggs pulls the house down," respectively) involve sequences of very long shots, which are exacerbated by degradation of shot detection due to poor lighting. In both cases the diminishing contribution of shot length to pace has been reflected by the new normalization scheme, resulting in a more intuitive drop in the pace level for each scene. Analysis of the differences caused by the new scheme is looked at in further detail below.

Fig. 6 and Table III show new tempo results from LW2 for shots 450–850. Overall, the new scheme resulted in a number of useful edges emerging (or, in some cases, being made more pronounced). As an example, it can be seen that the section between the end of the fight at the crooks house (edge A) and the start of the ensuing car chase (edge B) has been more accurately resolved in the bottom plot. This is a result of the increased sensitivity to the rise in shot length just above the median. Previously the large amount of camera motion in that area caused the tempo of the linking shots to be smoothed, resulting in no clear edges, as seen in the top plot of Fig. 6.

VI. Discussion

The pace function is a simple measure of a high-level construct compared to many of the features referred to in the background section. The power of $P(n)$ comes from, and confirms the usefulness of, an approach that is founded in a consideration of the forces at work in film construction. This power is evident from the fact that, in this experiment and others like it (of a less formal nature), it is able to capture information (e.g., scene breaks) without recourse to highly complex or computationally expensive solutions. In addition to this, the tempo measure avoids the problem of a fragile domain by virtue of it being closely tied to fundmental (physiological) human responses.

Taking the original pace function for a basis, we undertook two further investigations, namely, the improvement of the shot length contribution component, and the variation of the relative component weights. In the after analysis these investigations manifest themselves as being of two distinct natures.

The shot length normalization improvements are a refinement to the general pace function. Refinement to the pace function, where possible, is desirable as it is a foundational component of film construction and consequent appreciation. It is imperative therefore that it be hardened to the greatest degree possible, if it is to serve as the base for further and increasingly complex or less understood "semantic buildings." This is achieved by

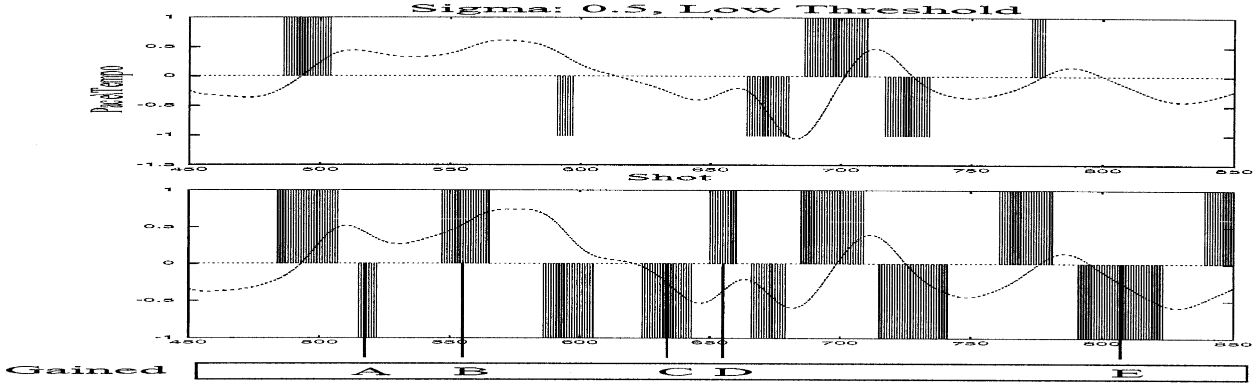

Fig. 6. *Lethal Weapon 2*—Top: original pace plot. Bottom: pace plot and edges gained with the new shot length normalization scheme.

TABLE III
NEW EDGES (SEE Fig. 6) AND CORRESPONDING STORY SECTIONS IN LW2 WITH THE NEW SHOT LENGTH NORMALIZATION SCHEME IN TEMPO COMPUTATION

	Edges Gained
A	End of fight, crooks house
B	Start of car chase
C	Boss arrives
D	Police leave crooks house
E	Riggs left with Murtaugh on bomb

reflecting as much knowledge about pace in our measure as is feasible and general, as gleaned from the film literature. The increased faithfulness to the actual pace is not as obvious from the outcome of the edge experiment, but an experiment focused on questions such as relative pace levels within or across movies would better reflect these improvements.

The component weighting work is concerned with exploring what happens when the pace function is attuned off center (i.e., is biased). It is effective at targeting different directorial and film styles and situations. As such it is not a refinement to $P(n)$, but an example of how the generic function can be sensitized to conditions of special interest. Such conditions, including any of those outlined in the results of the experiments itself, could be based on *a priori* knowledge regarding genre, could be directed manually via query relevance feedback, or be presented as a suite of results, either as a consolidated whole or combined with heuristics for the purpose of estimating likely content (e.g., large overall motion statistics might see a shift weighting shift toward the motion component for better sequence resolution).

It should be noted that generally *film grammar* does not codify content. For example, a rising shot rate does not equate to a car chase. What it does indicate, in this case, is that the director is doing *something*; raising the pace level, heightening demands on the audience, and this for a purpose. What that something is can be further fenced by means of existing techniques (e.g., a change of overall color composition might lend weight to the hypothesis that the shot rate change is accompanying an event that has caused a setting transition) but the level of interpreting signs with near 100% accuracy in film in the general case is beyond automation. This is due primarily to the nature of the film language. Rather than being a set of one-to-one mappings of signifier to signified, it is closer to a multilayered many-to-many language. ([33], [5])

There remains further scope for improvement of the pace measure, both in its essence and execution. We have completed an implementation of $P(n)$ that includes sound level and it already offers valuable input. Other factors like dialogue or visual disparity (i.e., setting and character turnover leading to increased demand on the viewer resources) could be examined in detail and assessed as to the likelihood of their being successfully integrated with the measure. Computationally, the desirable trait of real-time execution could be realized with the addition of statistical heuristics and iterative summary approximation. Such performance is valuable, particularly when seeking to farm out processing load, or in the absence of sufficient persistent storage or film source availability (i.e., streaming).

Consideration of an immediate application of a new result and a new technology is often illuminating. Such a tool as has been here discussed has potential both for content understanding and creation. An application for automatic content understanding has been demonstrated in the event/section experiments detailed here. Such tools would be useful in all manner of database annotation query and retrieval, on the fly trailer/synopsis for VOD-like systems, etc. Connected with this is content analysis at the time of creation or repurposing, where, for example, a filmmaker might use such tools to examine and reshape their work in an iterative fashion. From the perspective of pure content creation, particularly for the growing body of amateur filmmakers, such tools might effectively provide the function of a virtual director at the desktop. The effect here would be to distill the body of film literature into tangible aids.

VII. CONCLUSION

In seeking to create tools for the automatic understanding of film, we have stated the problem as one of faithfully reflecting the forces at play in film construction; that is, to interpret the data with its maker's eye. We defined *film grammar*, the body of film literature that effectively defines this "eye," and proceeded to take an example of carrying one aspect of this grammar from literature to computable entity, namely, movie tempo.

Having developed the tempo/pace function based on its filmic definition, we proceeded to examine the ends to which directors manipulate it. An application of this study was undertaken with the goal being to automatically locate dramatic events and section boundaries, the results of which were the successful reconstruction of the dramatic development of a number of films.

The above example not withstanding, the emphasis of this paper, and our central contribution is to draw attention to the realities of the film creation process. In particular, the importance of seeking meaning (semantics) in film from the source, namely, the filmmaker(s). Grappling with *film grammar*, actually and systematically, is critical to the problem.

In line with that thrust, the backbone of our future work will entail the further fleshing out of our understanding of *film grammar*. This will yield new aspects to be calculated, and improvements to existing measures. Our work on film rhythm [34] is completed, and performs well while being computationally tractable. Other targets include tone and mood, as well as certain "content" issues such as comparative genre studies.

It should be noted, in closing, that film is not the only domain with a grammar to be exploited. News, sitcoms, sports, etc. all have more or less complex grammars that may be used to capture their crafted structure.

References

[1] A. Smeulders, M. Worring, S. Santini, and A. Gupta, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 1349–1380, Dec. 2000.

[2] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: The QBIC system," in *Intelligent Multimedia Information Retrieval*, M. T. Maybury, Ed. Cambridge, MA: MIT Press, 1997, pp. 7–22.

[3] A. Pentland, R. Picard, and S. Sclaroff, "Photobook: Tools for content-based manipulation of image databases," in *Proc. SPIE, Storage and Retrieval of Image and Video Databases II*, 1994, pp. 2185–2205.

[4] D. Arijon, *Grammar of the Film Language*. Los Angeles, CA: Silman-James Press, 1976.

[5] J. Monaco, *How to Read a Film: The Art, Technology, Language, History and Theory of Film and Media*. London, U.K.: Oxford Univ. Press, 1981.

[6] T. Sobchack and V. Sobchack, *An Introduction to Film*. Glenview, IL: Scott, Foresman, 1987.

[7] B. Adams, C. Dorai, and S. Venkatesh, "Toward automatic extraction of expressive elements from motion pictures: Tempo," in *Proc. IEEE ICME*, vol. II, July 2000, pp. 641–645.

[8] ——, "Study of shot length and motion as contributing factors to movie tempo," in *Proc. 8th ACM ICM*, Nov. 2000, pp. 353–355.

[9] ——, "Role of shot length in characterizing tempo and dramatic story sections in motion pictures," in *IEEE PRCM*, Dec. 2000, pp. 54–57.

[10] R. Lienhart, S. Pfeiffer, and W. Effelsberg, "Video abstracting," *Commun. ACM*, vol. 40, no. 12, pp. 54–63, 1997.

[11] D. Bordwell and K. Thompson, *Film Art*, 5th ed. New York: McGraw-Hill, 1997.

[12] S. Fischer, R. Lienhart, and W. Effelsberg, "Automatic recognition of film genres," Univ. Mannheim, Mannheim, Germany, Tech. Rep., 1995.

[13] R. Hammoud, L. Chen, and D. Fontaine, "An extensible spatial-temporal model for semantic video segmentation," presented at the Proc. 1st Int. Forum Multimedia and Image Processing, Anchorage, Alaska, 1998.

[14] W. Mahdi, L. Chen, and D. Fontaine, "Improving the spatial-temporal clue based segmentation by the use of rhythm," presented at the Proc.ECDL, 1998.

[15] B. Salt, *Film Style and Technology: History and Analysis*. London, U.K.: Starword, 1992.

[16] N. Vasconcelos and A. Lippman, "Toward semantically meaningful feature spaces for the characterization of video content," presented at the Proc. ICIP, Santa Barbara, CA, 1997.

[17] ——, "Bayesian modeling of video editing and structure: Semantic features for video summarization and browsing," presented at the Proc. ICIP, Chicago, IL, 1998.

[18] A. Yoshitaka, T. Ishii, M. Hirakawa, and T. Ichikawa, "Content-based retrieval of video data by the grammar of film," presented at the IEEE Symp. Visual Languages, Capri, Italy, 1997.

[19] A. D. Doulamis, Y. S. Avrithis, N. D. Doulamis, and S. D. Kollias, "Interactive content-based retrieval in video databases using fuzzy classification and relevance feedback," presented at the Proc. ICMCS, 1999.

[20] H. Sundaram and S.-F. Chang, "Video scene segmentation using video and audio features," in *Proc. IEEE ICME*, New York, Aug. 2000, pp. 1145–1148.

[21] ——, "Determining computable scenes in films and their structures using audio-visual memory models," in *Proc. 8th ACM ICMM*, 2000, pp. 95–104.

[22] J. R. Kender and B.-L. Yeo, "Video scene segmentation via continuous video coherence," IBM T. J. Watson Research Center, Yorktown Heights, NY, Tech. Rep., 1997.

[23] I. Radev, N. Pissinou, and K. Makki, "Film video modeling," in *Proc. 1999 Workshop KDEX*, Chicago, IL, Nov. 1999, pp. 122–128.

[24] J. S. Katz, *A Curriculum in Film*, Toronto, Canada: The Ontario Institute for Studies in Education, 1972.

[25] H. Zettl, *Sight Sound Motion Applied Media Aesthetics*. Belmont, CA: Wadsworth, 1998.

[26] M. V. Srinivasan, S. Venkatesh, and R. Hosie, "Qualitative extraction of camera parameters," *Pattern Recognit.*, vol. 30, no. 4, pp. 593–606, 1997.

[27] Mediaware solutions WebFlix Pro v1.5.3 (1999). [Online]. Available: http://www.mediaware.com.au/webflix.html

[28] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1988.

[29] R. Deriche, "Recursively implementing the Gaussian and its derivatives," in *Proc. 2nd Singapore Int. Conf. Image Processing*, 1992, pp. 263–267.

[30] N. Vasconcelos and A. Lippman, "A Bayesian video modeling framework for shot segmentation and content characterization," presented at the Proc. CVPR, San Juan, Puerto Rico, 1997.

[31] Traditional film editing versus electronic nonlinear film editing: A comparison of feature films (1998). [Online]. Available: http://www.nonlinear3.com/brandt.htm

[32] *NIST Dataplot Software*, 1999.

[33] J. L. Salvaggio, *A Theory of Film Language*: Arno Press, 1980.

[34] B. Adams, C. Dorai, and S. Venkatesh, "Automated film rhythm extraction for scene analysis," in *Proc. IEEE ICME*, Aug. 2001, pp. 1056–1059.

Brett Adams received the B.E. degree in information technology from the University of Western Australia, Perth, in 1995. He is currently pursuing the Ph.D. degree at the Curtin University of Technology, Perth.

He worked for three years developing software, particularly for the mining industry. His research interests include content retrieval systems and tools, with a particular emphasis on mining multimedia for meaning.

Chitra Dorai (S'89–M'96–SM'01) received the B.Tech. degree from the Indian Institute of Technology, Madras, the M.S. degree from the Indian Institute of Science, Bangalore, and the Ph.D. degree from the Department of Computer Science at Michigan State University, East Lansing.

Currently, she is a Research Staff Member with the Internet Infrastructure and Computing Utilities Department at the IBM T. J. Watson Research Center, Yorktown Heights, NY. Her research interests are in the areas of e-learning, multimedia systems and analysis, computer vision, pattern recognition, and machine learning.

Dr. Dorai received the Best Paper Prize at the IEEE Pacific-Rim Conference on Multimedia in December 2000, for her work on movie tempo analysis. She also won the Best Industrial-Related Paper Award at the International Conference on Pattern Recognition in August 1998, for her research on automatic video text extraction. Her *Pattern Recognition Journal* article in 1997 received Honorable Mention in the 24th Annual Best Paper Award Contest of the Pattern Recognition Society. She is a member of the ACM.

Svetha Venkatesh (M'82–SM'92) is currently a Professor with the School of Computing, Curtin University of Technology, Perth, Western Australia. Her research is in the areas of large-scale pattern recognition, image understanding, and applications of computer vision to image and video indexing and retrieval. She is author of approximately 200 research papers in these areas and is currently co-director for the Center of Excellence in Intelligent Operations Management.