RC22655 (W0211-277) November 27, 2002
Computer Science

IBM Research Report

Formulating Film Tempo

[*]Brett Adams, Chitra Dorai, [*]Svetha Venkatesh

IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

[*]Department of Computer Science
Curtin University of Technology
Perth, W. Australia

Research Division
Almaden - Austin - Beijing - Delhi - Haifa - India - T. J. Watson - Tokyo - Zurich

Chapter 4

FORMULATING FILM TEMPO

The Computational Media Aesthetics Methodology in Practice

Brett Adams
Department of Computer Science
Curtin University of Technology
Perth, W. Australia
adamsb@cs.curtin.edu.au

Chitra Dorai
IBM Thomas J. Watson Research Center
Yorktown Heights, New York, USA
dorai@watson.ibm.com

Svetha Venkatesh
Department of Computer Science
Curtin University of Technology
Perth, W. Australia
svetha@cs.curtin.edu.au

Abstract

This chapter constitutes a detailed example of Computational Media Aesthetics at work. A short history of approaches to the problems posed by automatic content management in its broadest context is presented, cast in the light of their ability to obtain the much needed semantic grid with which to interpret their object. Our consideration is then further narrowed to the video medium, where we identify two common categories of solution to the problem, the *kitchen sink* and the *brittle mapping*, note their relative weaknesses, and show them to be directly attributable to the nature of the semantic grid chosen, or lack thereof. Focusing on our domain of Film, we argue that the best semantic grid for its interpretation is that within which its creators work; namely *film grammar*. In order to demon-

strate this, we develop a measure for the extraction of a fundamental aspect of film, Tempo. From definition, to formulation, and even its exploitation resulting in the location of high-level filmic components such as dramatic occurrences, the process is guided by film grammar at every step. Example results are provided from the movie, The Matrix.

Keywords:

Computational Media Aesthetics, video archive, content management systems, content-based search and annotation, semantic gap, motion pictures, film grammar, expressive elements, tempo, film pace, subjective time, dramatic sections, events

Introduction

While the technology savvy pundits thrill to new vistas of the "multimedia presence" opened to them by the efficacies of a relentlessly progressing storage and transmission technology, and even those normally more reticent in the giving of their time and trust are dragged along in its wake, the research literature bulges with an increasing testimony to the inadequacy of content management tools that are vital to reaping the true benefits of the technology of this new era.

A book without hide nor hair of contents page or index is an annoyance; a data warehouse containing Terabytes of video is next to useless without a means of searching, browsing and indexing such a deluge of data. In short, such content is wasted without suitable content *management*.

This chapter will outline an example of the application of the framework of *Computational Media Aesthetics* to the problem of automated understanding of film for media archiving and indexing. It seeks to leverage *film grammar* as the means for closing the semantic gap, the catch cry of today's research literature, and more importantly, the deficiency reliably identified as the true crux of the problem in today's content management systems.

The layout for the rest of this chapter will be as follows. Section 1 will focus on the current state of quandary into which the research field of multimedia content management and its various limbs, has stumbled, and an analysis of the origins of this situation. The implications of this in terms of a solution are drawn out for the wider domain of media content in general, and more specifically for the domain of film, thus positioning our work. The framework of Computational Media Aesthetics, in particular the use of the grammar of film, is presented as a solution, and its substance is presented in detail in Section 2. In Section 3 the reader is brought to the cutting face of the problem of film content management, in order to demonstrate the workings of the framework in detail, as it is applied to the problem of extracting tempo from motion pic-

tures. This includes the definition, formulation and extraction of tempo in line with its presence in the film literature. Finally, a conclusion is drawn regarding the usefulness of this approach, and its potential that may be exploited, or problems that must be overcome, in the future.

1. The Need for a Framework: Computational Media Aesthetics

The problem alluded to above can be summarized simply as: the obscurity of multimedia data to traditional techniques of information mining. To grasp why this is the case a short history of approaches to the problem is in order.

1.1 A Short History of Automatic Content Management

First, let us consider the problem of multimedia content management in its broadest context. Advances in technology have seen the area expand and successfully take in the domains of text, image, sound, and video, bringing us into the present. With the advent of the feasibility of large image databases/stores, the first idea for the task of managing that visual content was obviously to recourse to the mature technologies that had proven so successful. An example of the crudest form of this is that of annotating the image with a textual description, followed by simple indexing on that text with whatever embellishments are desired/possible (e.g., thesauri, related topic lookup, etc.). [26] has an example of a simple document classification system that uses Bayesian techniques with simple assumptions to provide a surprisingly good result, and serves to illustrate the types of textual "features" that may be leveraged to good effect (e.g., word counts, proximities etc.). However, the problem remains for our image query. Textual annotation is both labour intensive, and subject to the irregularities that arise from subjective labelling, and therefore a better solution was sought.

Instead of manually attaching what amounts to a narrow, cross-medium view on the image, why not find the word equivalent of the image itself? If, after all, a picture paints a thousand words, then it is sufficient to find these visual words, and immediately the existing techniques can be brought wholesale into this new medium. This is essentially what occurred with systems such as QBIC and Photobook [15] [29]. The image-centric words that were extracted were color, texture, shape and so forth. This was a useful result, but still the resulting user experience can be frustrating for certain large query domains, and for those of us without enough knowledge of the system to quickly hone our queries. Why is this? Hold that question as we move forward into the domain of video.

With another quantum leap in storage, and transmission technologies, compression techniques etc., we have the world of video arrive; the combined modalities of sound and images in time. The general movement was again to reharness the tackle of the traditional framework to this new horse. It was quickly apparent, however, that the existing paradigm would fit even more poorly with video, as it is a horse of a different color, and worse, it has wings. What constitute the "words" of a visual/audio stream? A typical video contains images at the rate of 25 per second. That amounts to something of the order of 135000 images for an average length film. Add to this the audio track, which is perhaps sampled at 44.1kHz, possibly in stereo. What features are pertinent amidst this tidal wave of data? In an ideal world you find them all and save those deemed pertinent, but the world is never ideal, and resource limitations always demand decisions from any would be classifier. We will examine this in more detail below, but suffice it to say here that the results are even less convincing than those for the image medium. For now, though, let us dig deeper to see if we can uncover the root of this lack of effectiveness of the re-tooling of methods tried and true.

Returning to the image domain we find that [33] has made a useful critique. Santini states that "images are not predicates, and can enter a relation of signification only when included in an externally defined discourse". Put simply, what this means is that without a *schema* directing and constraining the field of meanings possible from an image, semantic inferences cannot be made. This casts light on the efficacy of the textual query. A schema determining the interpretive grid through which the text must be understood is exactly what a typical text database supplies. For the image database, [33] asserts the existence of at least 3 semantic modalities (*linguistic*, *closed world* and *user*), and that the particular modality in force must be recognized before useful interaction can take place. For example, medical images of a certain type bring with them a distinct, shared interpretive grid through which inferences such as "healthy" or "diseased" can be made. In contrast, the significance of an image steeped in white might depend on a cultural context (e.g., western or eastern), or worse, on the specific context which becomes increasingly evident upon a user's interaction with the database.

With this insight, we see now why the problem becomes that much more pronounced as we move into the video domain. Each mode, aural and visual, not only brings its own unique context ambiguities, but a profusion of complex interrelations that result in a whole that is greater than the sum of the parts (e.g., rhythmic contrapunction).

Is the textual query mature? If that question refers to the domain of databases with well specified schema, then yes. But, for a burgeoning field such as natural language understanding, and the applications that hang from it, we must be a little more reserved.

1.2 Approaches to Film Content Management

As it is, we propose to expound the advantages of Computational Media Aesthetics [12] for the purpose of characterizing and managing film content, let us now zoom in on this chosen domain and take a look at the problem in action. We have stated that it is the want of a sufficient interpretive grid through which meaning may be derived from film data that has lead to difficulty with the extraction of semantic features. With the benefit of hindsight, we can see that the approaches have necessarily fallen into one of two categories.

- The first of these categories could be termed the *kitchen sink* approach. This solution is marked by classifiers/segmenters that integrate a profusion of audio-visual features. Crudely, this approach in effect says "lacking the necessary guidance of a grid, we couch similarity in terms that exist in our stable of features, and the more the better". The result is often that the technology drives the grid, as the criterion for the inclusion of feature X is simply that it is computable.

  This approach naturally gravitates toward comparing sections of data, with the consequent application of finding similar chunks and boundaries between them. Aside from the fact that film grammar asserts that the oft sought scene/sequence structure is little codified, the real problem here is that in actuality the semantic similarity of two pieces of film does not always coincide with the set of features which are used to infer similarity. Often the elements that bind a series of shots into a meaningful "sequence" are not color or motion, which are staples of these solutions, but music, a character, a theme, or any of a myriad of cinematic devices that are available to the filmmakers for the purpose of reinforcing this unity.

  Examples of this methodology include [21, 13, 20, 6, 19, 5, 44, 17]. Many of these solutions are augmented with temporal constraints, springing from a variety of models with an equal variety of implementations, depending on the precise application to which they are aimed. While some of these works are demonstrated in a proof of concept light, and practical considerations dictate that only a subset of the profusion of features that we mention are included, nevertheless, the stated or otherwise assumption remains that says, 'add enough of these features into the mix and you will find the effectiveness of the combined similarity metric increase'. The nub of the problem, however, is mentioned by [39], where they state that no one set of features is able to segment an entire movie. In other words, no one set of features is able to capture the entire breadth of expressive techniques open to the filmmaker for the purpose of crafting those semantically unified structures in a manner that

is also appropriate to the film's aesthetic demands. Their solution [39] is to throw the problem back into an interactive setting, and allow the user to dictate which set of features should come into play and when.

In order to detail one example of this, we will look at the contribution of [21]. There are two distinct parts to their solution. The first involves a measure of shot similarity, which is defined as being based on a mixing pot of "visual primitives such as colors, image correlation, optic flow, 3D hints etc.". A threshold is then applied to this similarity measure to form clusters of "semantically" related shots. These clusters are then subject to temporal constraints based on Allen's relations [7]. An underpinning of the resulting clustering is the assumption that transition effects (e.g., dissolve) occur at the boundaries of sequences, but film grammar does not support such an assumption in these absolute terms. A simple failure mode for this approach would be a sequence of shots that was bound solely on the presence of a character, or worse, a musical theme.

- The other category may be called the *brittle mapping* approach. The response here to the lack of a global grid is to set course amidst a sea of confusion for local landmarks that can be well defined/modeled. For example, if it is green and it has prickles it is a cactus. This approach gravitates toward classifying the nature of particular portions of a film's content, and hence constrains the query scope to exactly those modeled classes, lest the carefully crafted low-level to high-level feature mapping breaks down.

  Examples of this approach include [42, 38, 14, 28]. The down side of this approach is that either query scope or domain must be limited to a severe degree, otherwise the mappings breakdown completely. In the terms that we have been using, with the absence of a global interpretive grid, these approaches carve for themselves a local grid through which they can make inferences about highly specific objects.

  A good example of this approach is provided by [38], who aim to "establish a map between semantic attributes and measurable image features". They employ a Bayesian net for the purpose of inferring 4 particular types of *content* – Action, Close-Up, Crowd, and Setting – from the presence of 3 primitive features, motion energy, skin color, and texture energy. As an extreme example of a failure mode of this approach, the mapping between skin color and the presence of human(s) would show its brittleness if the human characters were painted an unusual color (e.g., such as many Sci-fi movies/TV series).

In summary, on the one hand, the kitchen sink gives us applicability to a broad domain, but with an accordingly higher degree of failure, whilst on the

other, brittle mappings give us reasonable performance with the price of a domain of practice that can be often limited beyond usefulness. How are we going to get the reliability of one, with the usefulness of the other? We believe that film grammar is the antidote to both of these problems. It constitutes the global grid which enumerates the variety of ways in which a sequence can be bound together for category (1), and exposes the expressive techniques that are both fundamental to the medium, and hence always present, and strongly tied to purposeful semantic ends for category (2).

2. The Solution: The Framework of Film Grammar

To overstate the point then: We need a better grid. Google (*www.google.com*) entered the web search game with *PageRank*, thus aligning its understanding of the data with the grid of user understanding. Santini [33] makes explicit three semantic modalities that must be recognized before the image search can effectively move forward.

How do we deal with the film domain? To define one's domain is no less than identifying (or at least, constraining) the candidate forces responsible for the data under investigation. The very same data can lead to wildly diverging inferences depending on the domain that is considered responsible for it.

2.1 What is Film Grammar?

Our chosen domain is film and the most complete source of information regarding the shaping forces of this domain is found in *film grammar*.

> **DEFINITION**: Film Grammar, is comprised of the body of "rules" and conventions that are a product of experimentation, an accumulation of solutions found by everyday practice of the craft [8, p. 2], [27, p. 119], [34, p. 189].

It elucidates the relationships between cinematic elements employed by filmmakers to convey meaning. The existence of the grammar makes obvious a vital truth, often overlooked, that films are crafted, built, and shaped with a purpose, regardless of what that purpose is. Given that films are objects whose construction is the result of intelligent consideration, that done by the film's collective "author"[1], it would seem that the obvious starting place for an interpretive grid for such a thing is indeed with that author and his intentions (see [18] for parallels in traditional text interpretation. Hirsch makes the point that the only universally valid interpretation of a text is that which the author intended. Salt [31] makes the charge that it is precisely because this point has been ignored that the field of film theory has drifted so far from what he terms

are scientific grounds for discourse). The problem becomes one of thinking the filmmaker's thoughts after him.

There are many and varied treatises on the exact nature of film grammar/-syntax, and the question of in what way it constitutes or contributes to a language of film ([32, 27, 25]), but there are some aspects of it that are generally agreed upon which are important for us to note if we are to handle it correctly. The chief among these should be obvious given the above definition, namely that it is descriptive rather than prescriptive in nature. It has grown up over the decades of filmmaking, and has formed in pragmatic airs; this has been found to work and has been added to the growing corpus, whereas that did not and was discarded.

Before we go on to look at how we handle the grammar given this situation, we must first answer the question that is immediately raised at this point: If film grammar is largely constituted by convention, what's to stop filmmakers flouting convention en masse and thus rendering such a "grammar" useless? Vogler [41, p. xiii] makes the point that "these artists run the risk of reaching a limited audience, because most people can't relate to totally unconventional art. By definition it doesn't intersect with commonly held patterns of experience"[2].

2.2 How do We Use Film Grammar?

Objections noted, we can proceed to analyze this grammar with closer scrutiny. Film grammar covers a broad, complex spectrum of more or less tangible aspects. At one extreme are the explicit rules, e.g., don't cross the eyeline. In such cases the computational media aesthetist rubs his hands with glee and sets about seeing what he can make of this jewel. In the case of the eyeline rule, it could, for example, provide a guiding input to an attempt at quasi-spatial reconstruction of a series of shots.

Unfortunately statements in the imperative mood regarding fundamental aspects of the craft are fairly rare, a fact that led Salvaggio to explicitly enumerate some of those rules [32]. In the middle of the spectrum, then, is where the vast bulk of film grammar lies, and is where fishing is rewarded. Here you have the largest bulk of expressive techniques and rules of thumb that are used to effectively impact the audience and express the film's narrative. They are stated with more or less vagueness in what amounts sometimes to a curiously indiscriminate tarring of different concepts with the same brush. Often useful concepts will nestle in the niches between the written line. But the information is there nevertheless, and in a manner that rewards careful sifting and synthesizing. The occasional vagueness is no doubt a function of precisely the fact that most often the "rules" are "unwritten", and thus a standard nomenclature has yet to emerge, combined with the nature of the subject. For example,

much is made of the desire to find 'scene' boundaries. The term comes to us from Greek, where it referred to the background to a dramatic performance, roughly equivalent to *stage*, via Latin and then classical French (theater) where it formed part of the structure of the play itself. It had a precise beginning and ending corresponding to the arrival and departure of a character. With the advent and maturation of the film medium the term was naturally absorbed into the idiom, but with a scope widened in proportion to the opportunities available to the roving camera and other cinematic techniques. Unlike the stage, film can truly have multiple parallel streams of narrative intertwined spanning locale and time. Hence, when applied to modern filmmaking "the term *scene* is useful but not precise" [27, p. 130]. The term is notably absent from the production process, and Katz notes that "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*".

At the other end of the spectrum we drift off into speculation and esoteric theorizing, and away from the "grammar" of film. This is the realm of Marxist theory, and psychoanalysis, and theories formed on the basis of a handful of films. Here, where we find very little consensus among the theoreticians, and a continually moving ground is no place to build a film schema. Salt [31], from his vantage point as a "Scientific Realis[t]", whilst not claiming that the study of film is a real science, nevertheless says that it will benefit from a scientific approach, and is particularly vitriolic with regard to the indulgences of the aforementioned theorists. And so, somewhat perplexed by all this, we leave it where it lies (and shifts), and return to the warmer waters described above.

In our sifting, rewards are had by considering the complete production process from first to last; screenwriter, set designer, director, cinematographer, editor, composer, etc., each has their own unique influence on the construction of the film. [16] explains a typical Hollywood style movie life cycle. The story, possibly a novel, possibly an idea, possibly a drafted screenplay, is given the go ahead through the agency of a producer, star, studio executive's wife's brother's neighbor; a director is brought on board; a shooting script is hammered out amidst a plethora of considerations (logistical and personal), another is made, possibly by the original screenwriter; stars are hired, sets are designed, shooting is done, as best as can be done; finals are edited, music is scored or ripped from the director's favorite CD, marketeers market; and finally the film *exists* in canonical form, or possibly several *canonical* forms. Obviously this is not the model for films of every persuasion and nationality, but it does serve to illustrate some of the creative forces that come to bear upon it between inception and completion. And let us not forget the storehouse of insight that is common sense and experience. We, the armchair film critics, are also able to make useful contributions to the recognition of important elements of film.

What are some of the key attributes to look for in a candidate filmic element? Practically speaking, the most initial gains will be had if we choose elements that have shown longevity and broad scope[3]. What change, if any, has it demonstrated over time and why? Is it a passing fad? Or does it have deeper roots in the human/media dynamic experience? Is it something more fundamental to the medium? Insight into these questions may often be obtained by examining the historical reasons for it's inclusion in the corpus. E.g., was the origin of the *cut* due to technological or artistic considerations, or a combination of both? Some of these questions may benefit from a consideration of the extent to which the element taps the firmer ground of underlying human response to certain stimuli.

Dorai and Venkatesh [12] have recently placed such considerations in media analysis on the firm foundations of film grammar, by proposing an approach called the Computational Media Aesthetics. It is the algorithmic study of a variety of image and aural elements in media, founded on their patterns of use in film grammar, and the computational analysis of the principles that have emerged underlying their manipulation, individually or jointly, in the creative art of clarifying, intensifying, and interpreting some event for the audience [12]. Computational Media Aesthetics provides a handle on interpreting and evaluating media elements in productions through knowledge of film codes that mediate perception, appreciation and sometimes rejection. It exposes the semantic and semiotic information embedded in media productions by focusing not merely on the representation of perceived content in digital video, but on the semantic connections between the cinematic elements in the content and their emotional, visceral appeal. It studies mappings between specific narrative forms and their intended affect.

Here ends the general introduction to film grammar and Computational Media Aesthetics, and the means to its exploitation for the purpose of automatically understanding film. We will now detail a case study, as it were, of its use for the extraction of one particular aspect of film present in the tenets of film grammar, namely *tempo*.

3. Using the Framework: Extracting and Analyzing Film Tempo

One aspect of film grammar that is reliably referred to across the breadth of the film literature is what we will happily call *Tempo/Pace* now, and qualify shortly. The term tempo is already the result of a process of investigation and synthesis from the source literature.

3.1 What is Tempo?

An example definition is given by Zettl [43], who speaks of "Pace", which "refers to the perceived speed of an event", and is "felt time", or "subjective time". At the very least we can extract from this the fact that pace/tempo has to do with speed.

Propositional definitions, however, are rare. More often tempo is merely referred to, with accompanying descriptions, which the computational aesthetist will mine for whatever value is possible. For example, Sobchack [34] speaks of "*Perception of time*, which is subjective", and "might be thought to all a matter of opinion, or of feeling, but in fact most members of the audience will probably agree that a given film is light or swift or steady." Thus we see that tempo can be qualitatively described as swift or steady, frantic or slow [40, p. 205]. Does the film drag or move briskly?

Finally, many texts just casually refer to tempo, assuming that the reader understands what is meant by the term.

[40, p. 221] notes that "chases may occur in any part of the story, but the end of act two is one of the most popular places. Chases are useful for torquing up a story's energy ... a time when you want to pick up the *pace* and build momentum for the finish". McKee [22, p. 220] speaks of "accelerating the mid-film *pace*", while Mehring [24, 155] tries to confuse matters by using two tempo-like terms in the one sentence: "Through *tempo*, rhythm, and/or *pace* the screenwriter blends how something is said with what is said to create the desired viewer's response".

At this point of our investigation it is worth looking at the possible underlying factors of this concept to see if we can gain a fuller understanding of it. Experience supports the notions that are detailed in the literature snippets above. We know that movies ebb and flow, and this seems to happen as the filmmakers respectively hold and overwhelm us with information or activity. Reflecting on this, and taking some cues from the realm of psychology (e.g., [9] on the concept of perception of time, and the apprehension of the "psychological present") we might formalize the concept of tempo as being a function of the rate of delivery of information thrust at the viewer. This is not, by any means, sufficient grounds upon which to conclude the case for tempo, but is a worthwhile activity for the questions it raises. Returning to the film literature, we find that in the context of a passage concerned with tempo, Mehring [24, p. 146] relates boredom, or slow tempo, with "new shots that do not contain new information". With this perspective gained we see the idea of tempo being related to the flow of information in many places.

We have thus sharpened our definition of tempo to its quintessence, namely *rate of delivery* (from an objective point of view) with the desired goal being the manipulation of the audience perception of time. Intuitively this makes

sense, and brings to mind musical tempo (the rate at which a piece is to be played, e.g., allegro), alongside which our tempo sits as a filmic counterpart. From the variety of terms employed in the literature (including tempo, pace, subjective or felt time, perception of time, etc.), we choose tempo for simplicity sake, and aware of the fact that some authors, e.g., Zettl, do make a conscious, subtle distinction between *tempo* and *pace*, we nevertheless allow that they are intimately related, and for our purposes can be put together.

The question still remains: Is tempo worth extracting? We have seen that it has a large presence in the film literature, but does it exist in a suitably stable state to make its extraction worthwhile? Further to this, is it useful? In answer to the first question, we can see that the concept of film tempo has been recognized for a considerable time. Given the fact that we the audience are trained to watch movies (see [27, p. 121]), history has seen an inevitable desensitizing – what was once fast is not so fast anymore, as Sobchack notes of the Beatles "A Hard Day's Night": "When the film was released it left audiences exhilarated and breathless. Today, however, A Hard Day's Night seems much slower...the pace seems almost sluggish". But, this is an issue of comparative film analysis and degree, and can be addressed. The key point to note is that the presence of tempo remains, and it is indeed a fundamental component of film. The second question, regarding the usefulness of tempo, is answered by the preceeding assertion. *Any* fundamental aspect of film is worth extracting, even those that do not immediately lend themselves to application. Often the usefulness of one element is only seen when put next to others at a later date, and we dare not starve ourselves from the outset of such valuable sources. Sobchack assures us above, though, that tempo is not a recondite concept, subject of the ruminations of an elite few, but is appreciated at some level even by the average film audience, and hence is immediately of interest to us.

3.2 Manipulation of Tempo

Having defined what film tempo *is*, we now seek *how* it is created, the nuts and bolts of how it is manipulated for intended effect in a movie. Again we turn to the literature for our answers:

Sobchack [34] says that "[Tempo] is usually created chiefly by the rhythm of editing and by the pace of motion within the frame". The Encyclopedia Britannica adds that tempo is influenced by "the actual movement and cuts within the film, by the accompanying music and by the content of the story". In Zettl's [43, p. 285] terminology, subjective time is dependent on *event density*, which can be manipulated by "primary, secondary and tertiary motions", which correspond to object motion, camera motion, and "motion" induced by the edit. The zoom is included in this list as being able to impart "high energy" [43, p. 294].

Given this list of contributing factors, a balance must now be struck regarding computability and relative usefulness. In other words, what can we compute, and what will give us the most return. The two most dominant factors appear to be motion (both object and camera) and shot rate, with sound (score, incident sound and dialog) as a close third. The remaining factors, and there are many, are less frequent (e.g., zoom) but nevertheless computable, or more obscure factors that appear beyond computation at this point (e.g., story).

The extent to which each of these *carriers* of tempo are employed will differ for filmmaker (style), genre (appropriateness and convention) and possibly culture (a combination of the previous factors, and others, e.g., technology available), but they nevertheless remain the dominant underpinnings to tempo, and should the filmmaker wish to manipulate tempo (and we have established that invariably they will), one or all of these elements must be crafted. Even an extreme example like Alfred Hitchcock's "The Rope", which is singularly ambitious in its lack of visible editing, must necessarily manipulate camera and object motion and sound volume or rate in order to alter tempo and consequently audience subjective time.

3.3 Computational Aspects of Tempo

Our task has now been sharpened to the point of requiring a measure of the object/camera motion within the frame, and an index of shot boundaries, for the movie under consideration.

3.3.1 Extracting the Components of Tempo. For motion estimation there are many algorithms that might be used, but we have chosen an algorithm [35] that captures the camera movement in terms of pan and tilt between two consecutive frames. Medium to large object motion is also captured in these terms. The algorithm is applied to an MPEG-1 stream, encoded at 25 frames per second. Estimates that fall outside 3.5 standard deviations from the motion magnitude mean are deemed anomalous and thrown away. The resulting series of motion estimates is smoothed with a Savitsky-Golay filter [30], for the purpose of dampening noise. The Savitsky-Golay filter is used since the camera/object motion being reconstructed is assumed to have non-zero 2nd order components.

Candidates for determining an index of *cuts* are also plentiful, and we have used both commercial software [23], and in-house solutions [36]. Complex shot transition devices such as the fade or the dissolve are not extracted given the poor precision and recall performance of the current state of the art, and their relative rarity when compared to the cut. The resultant shot index series undergoes a merging process whereby shots under a length threshold are deemed false positive, and are joined with the preceding shot. A typical threshold is 7 frames (which equates to about 280 ms at 25f/s), which is an extremely

rare length for a shot, but long enough to account for false positives due to problematic content such as close-up motion.

3.3.2 Formulating Tempo. Given the palette of these raw attributes, motion estimates and shot index, how best can we paint a picture of tempo? How can we assemble these parts into a representation of tempo, one that reflects a reliable expression of the filmmaker-manipulation/viewer-appreciation of tempo?

The first criterion is that our representation faithfully capture ebb and flow, as we have already noted. Film grammar uses binary/ordinal adjectives like brisk and dragging to describe tempo, but also indicates that there can be gradual transitions between those states and that it can accelerate and dampen. From experience we know this to be true, and that far from being able to classify particular parts of a movie in absolute terms such as "fast", we are much apt to use incremental qualifiers such as "faster" with confidence. Therefore our formulation of tempo must capture this aspect, i.e., it should be a *continuous* measure.

A continuous measure over what base though? It ebbs and flows over time, but what time index will be the most useful? The options are two: A strict time-based index such as frame or second, or a quasi-time base such as *shot*. We have chosen the second, shot, as it is initially simpler due to the fact that one of the tempo contributors to be used *is* an editing attribute, i.e., tied to the shot. In addition to this, using shot as our base does not prevent us from flattening the measure onto a pure time base when necessary, as we have a shot to time mapping in the form of the shot index.

Having identified the contributors to tempo that we wish to include in the measure, the next question is *how* are they to be included? Each component must make a contribution that is reflective of its part in manifesting the tempo of a film to the viewer.

First, let us look at motion. The link between motion and tempo should be sought in light of our statement concerning the *essence* of tempo, namely that it is a function of the rate of information being thrown at the viewer. What indications about information can motion estimates provide? Motion implies new content of some form has entered the field of view, whether it be objects entering the frame through their own motion, or by virtue of a moving camera and consequent changing field of view. The motion may not even bring *new* content, but it nevertheless will make demands of the viewer in terms of tracking with the changes and assimilating them (e.g., a swerving car, a capering comedian). Given this, we will make motion *magnitude* a proportional contribution to tempo, and with no impetus to proceed otherwise, we will normalize it in a standard fashion (see Equation 4.1).

The second component under consideration is editing. How does editing contribute to tempo? Specifically, what does a cut do to the rate of information presented to the viewer? References generally link tempo to the rate at which shots come, and therefore make tempo a function of shot length. Zettl [43] discusses the purposes and effects behind shot length manipulation in terms of event density and its effect on subjective time (tempo in our discussion). Thinking more deeply about why this is so, we find that film grammar also tells us that consecutive cuts to almost identical shots are a no no. All that such shots generally do is call attention to themselves, so we can assume that each cut introduces new information to be assimilated by the user, even if the demand is simply to recognize that there has been a cut back to a familiar character/pov/angle/etc. Even rapid cuts with little *new* information can torque tempo. [10] discusses the time required to adjust to a new shot, and uses the example of the strike on the terrorist camp in Patriot Games to demonstrate. Here we are battered with alternating shots of the satellite beamed images and the response of Jack Ryan, at an increasing rate that leaves you breathless because you have only begun to adjust to the new shot when the next comes crashing onto the screen. This means that cuts can be thought of as dropping *blobs* of information into the film stream that require resources of the viewer in order to assimilate this new information. Following a cut the viewer will either proceed to adjust to this new shot until either (a) the requirements placed on the viewer taper to zero when there is no new information to adjust to (in which case if the filmmaker does not want tempo to drop too much they must provide info in other modes - i.e., motion, dialog, story etc), or (b) a new cut is made which resets the amount of information to be assimilated, starting the process again.

It should be obvious that this cannot be modeled in the simple fashion with which we have dealt with motion. In order to faithfully capture the contribution of editing to tempo, we need to first consider what the director is trying to do with each *range* of shot lengths, find this spread across film grammar, and then to reflect that knowledge in a weighting function, the counterpart to the standard normalization used for motion.

We find that there is a hard lower limit to the shot length in practice, and this is due to ability of a human viewer to adjust to and process the information of a new shot. Except for the most extreme cases, it is desired that the viewer will receive and assimilate the information in a new shot and add it to the growing story. Often shots in this range will be part of a metric montage designed for percussive effect, the latter part of an increasing metric montage, or an isolated incidence calculated to shock. The aim with all of these is a fast tempo, and the shorter the shot, the faster the tempo.

Moving away from the lower limit, we come to the shot length median. Around it cluster shot lengths that are the result of normal narrative require-

ments. Rhythmic montage, medium paced metric montage, and many narrative restraints all result in shot lengths at or near the median. In a sense this is the default length for maintaining audience interest levels, all other contributions to this goal being equal. Tempo is still sensitive to adjustments made around this middle-point.

The upper limit of shot length appears to be rather amorphous. It is no longer physical limitations that dictate here, as did with the lower limit, but more complex factors such as, genre or target audience indicated conventions/wisdom as to what can be accepted, 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. Editing's contribution to tempo is attenuated as we move away from the median and come to this point, the burden of tempo creation must be shouldered by the other modes (i.e., motion, story, sound etc.) or else tempo drops to the overall minimum.

To summarize, tempo is most sensitive to shot length changes up to and near the median, but progressively less so as the length moves beyond this range to the maximum.

3.3.3 The Tempo Function. In light of the above analysis, the weighting function for shot length normalization is formulated thus: The median (which proves a better estimate of the average shot length in the presence of outliers) is chosen as the zero point of editing's contribution to tempo. Lengths below this point are well bounded by the minimum shot length below and are weighted with a simple linear model. The slope of the curve is calculated to weight the minimum shot length with a unit weighting. Above the median, where outliers can be far removed from the preponderance of lengths, a linear model would, in general, under-weight a large range of lengths. Instead, a curve of the form $x^{1/2}$ has been chosen. This results in sensitivity near the median, but a rapidly declining effect for lengths that are removed from the median. The curve is fitted to reach -1 at the 95^{th} percentile of shot lengths; -1 for symmetry, and asymptotically behaving at 95% to provide robustness against the farthest outliers. See Figure 4.1 for a plot of the shot length weighting function.

As an aside, it has been suggested (e.g., [37]) that a member of the Weibull family of distributions models shot length well. Our own empirical investigation supports this, with an loosely estimated alpha of 1.5, which means an increasing hazard function (interestingly, of the form $x^{1/2}$) (see Figure 4.2). In addition, this makes sense from a theoretical point of view. The Weibull distribution is often used to model the first failure from among a group of possible causes, which is certainly a valid way to view the process that leads to a cut. Each of a number of ever present factors may cause a cut, e.g., narrative

demands (new information beyond the field of view of the current shot must be displayed), or reaction shots for story elements (a piece of dialog), matters of tempo (the current shot has gone on too long), or even logistical considerations (the inability to continue a shot for technical reasons, or recourse to a *master shot* when an insert is deemed unsatisfactory, but unable to be reshot), the list goes on.

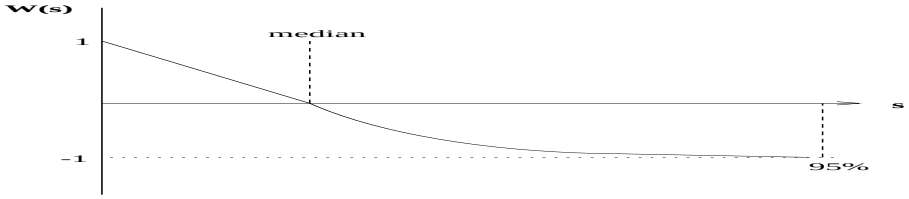
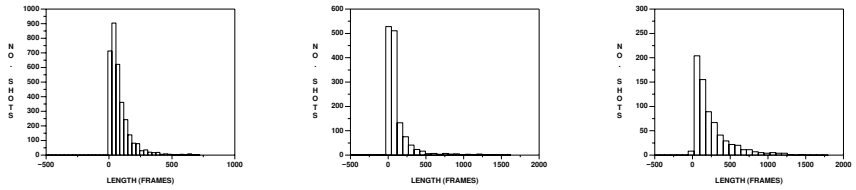


Figure 4.1. New shot length normalization scheme for tempo computation.



(a) (b) (c)

Figure 4.2. Shot length distributions: (a) Titanic; (b) Lethal Weapon 2; (c) Lost World Jurassic Park.

Now we can view the tempo function, T in its final form:

$$T(n) = \alpha(\, W(s(n))) + \frac{\beta(m(n) - \mu_m)}{\sigma_m}. \tag{4.1}$$

where s refers to shot length in frames, m to shot motion magnitude, and n to shot number. The standard deviation of motion is given by σ_m and μ_m is the mean of motion magnitude and these are used in normalizing motion estimates as shown. $W(s(n)$ is the weighting function of Figure 4.1 used in shot length normalization. The weights α and β, are given values of 1, effectively assuming that both shot length and motion contribute equally to the perception of tempo for a given film.

Integral to the finished function is the phase of Gaussian smoothing that is applied to the output of Equation 4.1. This smoothing is applied for two reasons:

1 It reflects knowledge of the domain, namely that filmmakers, excepting rare narrative requirements, do not make drastic, repeated tempo changes on the order of shots.

2 Zettl [43, p. 276] notes that "only a series of shots can have a pace[tempo]". That is, tempo has a certain inertia to it, and is therefore a function of a neighbourhood.

The amount of smoothing affects the resolution of the resulting tempo measure. The lower the amount, the more detailed are the features that can be located in the tempo landscape.

The interested reader is referred to [3], where we investigate other α and β weighting schemes, particularly with a view to their employment in crafting tempo in different movie genres. In short, the study examined the different roles of motion and shot length with regard to tempo in the action and drama genres. We found that greater sensitivity to shot length (α) resulted in increased resolution of the dramatic progression of both genres, while greater sensitivity to motion (β) saw increased resolution of action oriented events in the action genre, but no appreciable return from the drama genre.

3.3.4 An Example from the Movie, The Matrix. Now we are in a position to look at the result of the measure we have formulated for tempo. Figure 4.3 is a plot of the tempo for the movie "The Matrix". The shot length threshold used to find the shot index is 7 frames, α and β are set to 1, and a large amount of smoothing has been applied to the output (50 smoothing turns with a Gaussian of $\sigma = 1.5$), resulting in a low resolution curve useful for a quick overview of the movie. From this initial overview we get an idea of the ebb and flow of tempo occurring in The Matrix. We can immediately see that the movie has regularly placed tempo peaks that build in magnitude and breadth as the movie progresses, culminating in a prolonged period of high tempo, coincident with the movie's climax.

This sort of simple analysis is useful in its own right, and one might imagine a number of applications that could benefit from such information. The true worth of the measure, though, can be seen when we begin to use it as a base for deriving deeper inferences. What signatures in the measure might we look for and exploit for high level understanding? What composite features might tempo constitute but a part? Again, if the tempo measure provides reliable *information* about the given movie, then it will enable us to make true *statements* about the movie, which is precisely the enabling technology for a successful content management system. There remains, then, to be found what other true/reliable measures to which tempo can be coupled for correspondingly greater return. The next section will examine one such further use of tempo.

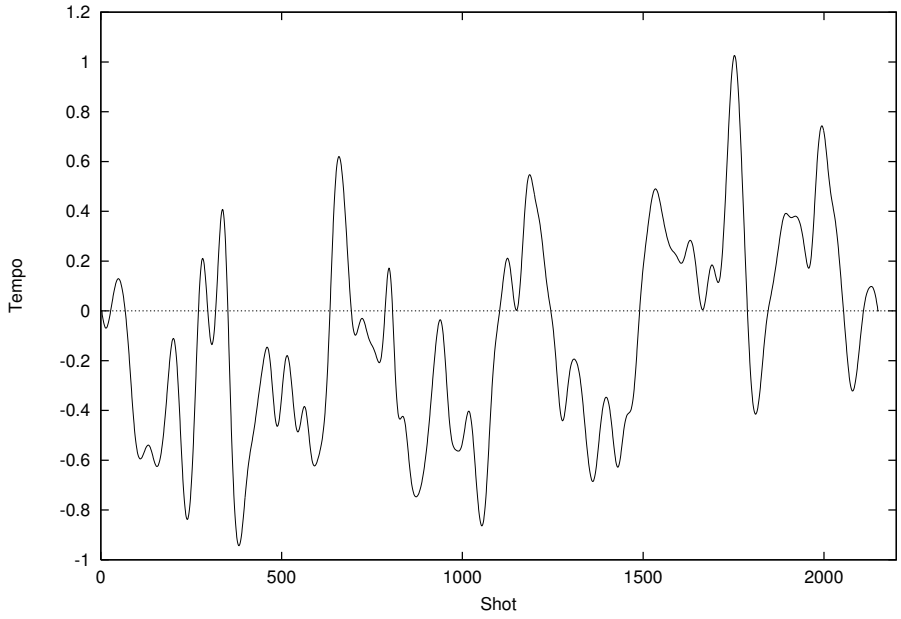


Figure 4.3. Tempo plot for the Matrix.

3.4 Building on the Tempo Function

Film grammar tells us that tempo changes often accompany story section boundaries and are precipitated by local events. By story section we mean that which is commonly termed a scene or sequence. It is a large piece of the narrative, and often is initiated by an abrupt jump to a new location, or perhaps an event with far reaching consequences. In the first case, a tempo change often serves to heighten and reinforce the sense that the story has indeed moved to a new location, encompassing a possibly vastly different set of characters or circumstances for the same characters. In the second case, a tempo change often issues from a powerful force that has been let loose and now inevitably moves the story into a different phase. The advent of the iceberg in Titanic, and the ensuing mayhem, torques the tempo of Titanic to a level that is never quite relaxed until the end of the movie (see [4] for the tempo plot of "Titanic"). [40, p 152] mentions that the "...audience will still experience a noticeable shift in energy at the threshold crossing. A song, a music cue or a drastic visual contrast may help signal the transition. The pace of the story may pick up".

Local events occur on a smaller scale, within a story section, and are the nuts and bolts of plot movement. They may be physical or emotional in nature,

but regardless of their form, they too result in incidents that inevitably must be captured by the filmmaker in film constructs that result in altered tempo, or are narrated that way in order to impact and alert the viewer to the import of this particular piece of the narrative, thus mirroring their larger cousins, the story section boundaries. [24, p 146] notes that "A sequence designed to communicate boredom suggests the monotony of measured shot lengths or new shots that do not contain new information. A race may call for progressively shorter shots [i.e., increasing tempo]."

Armed with this knowledge, we may have the ability to detect these film components. We also have the motivation to detect them, as at the very least their whereabouts would provide a valuable index into the dramatic progression of the movie in question.

Our task then is to detect tempo transitions. We approach the problem as one of detecting edges (transitions) in the tempo function. Edges of the tempo function are detected using Deriche's recursive filtering algorithm [11]. 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 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.

Two rounds of edge detection have been applied to the Matrix. *Large pace transitions* are targeted with a high threshold, and the resulting edges are designated *"Story sections"*. This label is somewhat arbitrary as large transitions do not always indicate a change of story section 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 the localized events that we have described above, as opposed to changes of the order of story section size. Figure 4.4 plots a portion of the results, and Table 4.1 supplies a manually supplied label of significance to each detected edge.

The reader is referred to [4, 3, 1, 2] for comprehensive edge analyses on a number of films.

So we can see that the detection of edge signatures has yielded further information to that which was immediately apparent from the raw tempo measure. The existing measure, coupled with fresh insights from film grammar, has provided this. This is offered as an exemplar of the approach that we are here advocating. That is, the cycle of gathering insight from film grammar, synthesizing it, building the computational equivalent where possible, returning to film grammar with sharpened sight, and the pooling of subsequent insights back into the growing technology base for content management.

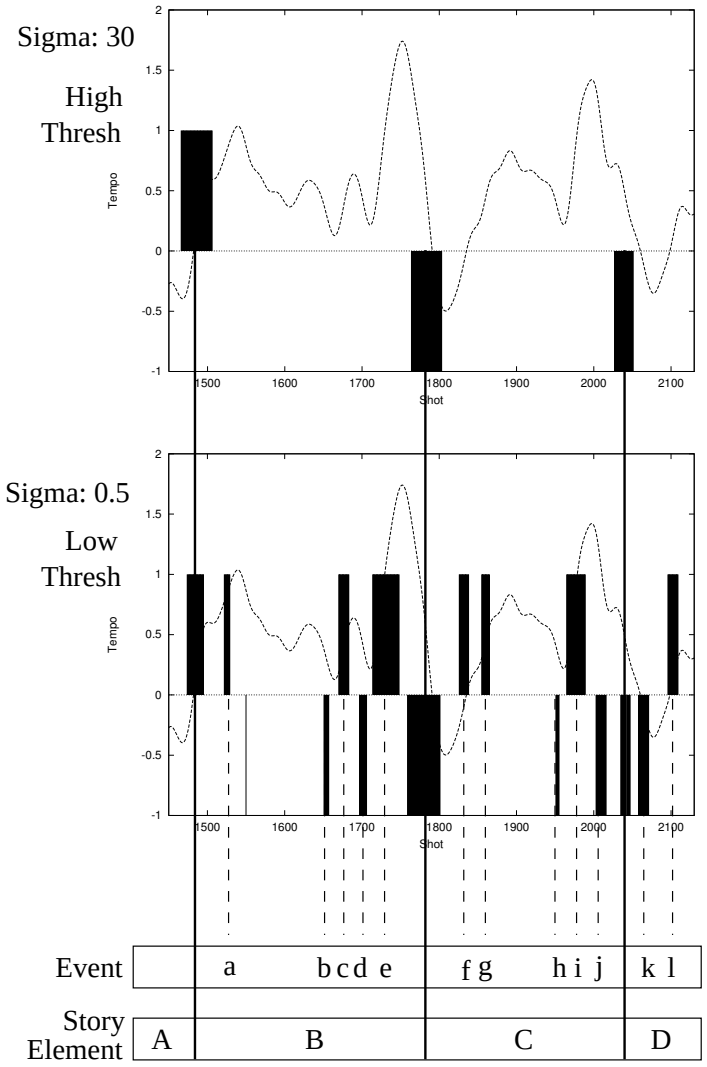


Figure 4.4. Results of edge detection on tempo flow and corresponding story sections and events from the Matrix.

Table 4.1. Labeled story sections and events identified from tempo changes in the Matrix (cf. Figure 4.4).

Story Section detected - High threshold	
A	Morphius in captivity
B	The rescue of Morphius
C	Neo finally confronts the matrix
D	Neo is the 'one'
Event detected - Low threshold	
a	Neo and Trinity face the soldiers
b	Trinity shoots the agent
c	The cavalry (helicopter) arrives to rescue Morphius
d	Morphius leaps to safety
e	The escape
f	An agent appears, Trinity escapes
g	Neo finally faces an agent
h	...and wins
i	Neo is chased
j	<False positive>
k	Neo is killed?
l	...No, he is alive, the climax

4. Conclusion

So where have we been? Starting with a short history of the field of automatic content management, we proceeded to highlight deficiencies in the current approaches, and postulated underlying causes for them. We grouped research in the area into two general categories, and then asserted that these diverging approaches result from different reactions to the lack of a semantic grid through which meaning may be attached to the data.

A solution was put forward, which for the domain of film constitutes that semantic grid, namely film grammar. Film grammar brings to the fore an oft forgotten fact that films are objects of sentient construction. They are crafted for a purpose, and as such are replete with the marks of their maker and take their form from the shaping of the elements of the grammar, harnessed to the narrative's end. We looked at how this grammar might be mined in general terms, including where the richest veins are to be found.

In order to demonstrate a concrete example of the process we then detailed our work on film tempo. We built upon this by using the tempo measure to find dramatic events and story section boundaries. This process of surveying

the film literature, to the final computational entities, is given as an exemplar of the Computational Media Aesthetics approach to the problem of automated media content management. Needless to say, the approach is not limited to film, but is applicable wherever similar grammars offer themselves as semantic grids, such as news, documentaries, sports, and the list goes on.

Where to from here? The attentive reader will have noticed that a number of elements that were isolated as being contributors to tempo were left, without further attention, in Section 3. One of these in particular is an extremely important element, namely sound. Without it, the tempo measure is in fact a subset of movie tempo; it is *visual tempo*. With this recognition, we have also undertaken work to include sound as a component of tempo in like manner to the other two components. This is especially important for movies that rely less on the manipulation of visual tempo, and more on aural cues (such as many films in the drama genre). Further work will include the identification and assimilation of zoom into the measure, as this too is a potent method of manipulating tempo.

The refinement of tempo is no doubt an endless road, with a precision that increases with diminishing returns, but is worthwhile nevertheless. Another profitable avenue for further exploration is illumined by questions like: "What filmic components are partly comprised of tempo itself?", or "What other high level constructions leave unmistakable signatures in the tempo landscape?". The answers are to be found in the same place.

Notes

1. Despite the fact that in many circles the concept of the auteur is of ill-repute, e.g., [16, p. 100].

2. Salvaggio [32] notes that "[The spectator] may not be able to formulate the rules, but he recognizes an ungrammatical sequence when he sees one." which answers another challenge that my be raised at this point, i.e., that one must be a film student to appreciate and make use of film structures.

3. [31] notes that his aim is "to produce new information about film style and its determinants that is not only true now, but will always be true to the greatest possible extent. This extent is limited by the essential uniqueness and idiosyncrasy of individual art objects, in our case films; and of course that idiosyncrasy and uniqueness is what makes some films art rather than craft. It is in the nature of artists to defeat the expectations of everyone next...".

References

[1] Adams, B., Dorai, C., and Venkatesh, S. Novel approach to determining movie tempo and dramatic story sections in motion pictures. In *2000 International Conference on Image Processing,* (September 2000), vol. II, pp. 283–286.

[2] Adams, B., Dorai, C., and Venkatesh, S. Role of shot length in characterizing tempo and dramatic story sections in motion pictures. In *IEEE Pacific Rim Conference on Multimedia 2000* (December 2000), pp. 54–57.

[3] Adams, B., Dorai, C., and Venkatesh, S. Study of shot length and motion as contributing factors to movie tempo. In *8th ACM International Conference on Multimedia* (November 2000), pp. 353–355.

[4] Adams, B., Dorai, C., and Venkatesh, S. Towards automatic extraction of expressive elements from motion pictures: Tempo. In *IEEE International Conference on Multimedia and Expo* (July 2000), vol. II, pp. 641–645.

[5] Adjeroh, D., Lee, M., and King, I. A distance measure for video sequences. *Computer Vision and Image Understanding 75*, 1/2 (July 1999), 25–45.

[6] Aigrain, P., Joly, P., and Longueville, V. Medium knowledge-based macro-segmentation of video into sequences. In *Mark Maybury, editor, Proc. IJCAI Workshop on Intelligent Multimedia Information Retrieval* (Montreal, August 1995), vol. 12.

[7] Allen, J.F. Maintaining knowledge about temporal intervals. *Communications of the Association for Computing Machinery 26*, 11 (1983), 832–843.

[8] Arijon, D. *Grammar of the Film Language*. Silman-James Press, 1976.

[9] Block, R.A. *Cognitive Models of Psychological Time*. Lawrence Erlbaum Associates, Publishers, 1990.

[10] Brandt, M. Traditional film editing vs. electronic nonlinear film editing: A comparison of feature films. http://www.nonlinear3.com/brandt.htm, 1998.

[11] Deriche, R. Recursively implementing the Gaussian and its derivatives. In *ICIP'92, Proc. 2nd Singapore Int. Conf. on Image Processing* (1992), pp. 263–267.

[12] Dorai, C. and Venkatesh, S. Computational Media Aesthetics: Finding meaning beautiful. *IEEE Multimedia 8*, 4 (October-December 2001), 10–12.

[13] Doulamis, A., Avrithis, Y., Doulamis, N., and Kollias, S. Interactive content-based retrieval in video databases using fuzzy classification and relevance feedback. In *ICMCS'97* (1999).

[14] Fischer, S., Lienhart, R., and Effelsberg, W. Automatic recognition of film genres. Tech. rep., University of Mannheim, Germany, 1995.

[15] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., and Yanker, P. Query by image and video content: The QBIC system. In *Intelligent multimedia information retrieval*, M. Maybury, Ed. MIT Press, 1997, pp. 7–22.

[16] Goldman, W. *Adventures in the Screen Trade: A Personal View of Hollywood*. Abacus, 2000.

[17] Hanjalic, A., Lagendijk, R., and Biemond, J. Automatically segmenting movies into logical story units. In *Visual Information and Information Systems* (1999), pp. 229–236.

[18] Hirsch, E.D. Jr. *Validity in Interpretation*. Yale University Press, 1967.

[19] Li, Y., Ming, W., and Kuo, C.-C. Semantic video content abstraction based on multiple cues. In *IEEE International Conference on Multimedia and Expo 2001, ICME2001* (Tokyo, Japan, August 2001), pp. 804–807.

[20] Lienhart, R., Pfeiffer, S., and Effelsberg, W. Video abstracting. *Communications of the ACM 40*, 12 (1997), 54–63.

[21] Mahdi, W., Chen, L., and Fontaine, D. Improving the spatial-temporal clue based segmentation by the use of rhythm. In *Second European Conference, ECDL '98* (1998).

[22] McKee, R. *Story: Substance, Structure, Style, and the Principles of Screenwriting*. ReganBooks, 1997.

[23] Mediaware Solutions Pty Ltd. Mediaware solutions WebFlix Pro v1.5.3. http://www.mediaware.com.au/webflix.html, 1999.

[24] Mehring, M. *The Screenplay: A Blend of Film Form and Content*. Focal Press, 1990.

[25] Metz, C. *Film Language: A Semiotics of the Cinema*. Oxford University Press, 1974.

[26] Mitchell, T.M. *Machine Learning*. New York, McGraw-Hill, 1997.

[27] Monaco, J. *How to Read a Film: The Art, Technology, Language, History and Theory of Film and Media*. Oxford University Press, 1981.

[28] Naphade, M. and Huang, T.S. A probabilistic framework for semantic video indexing, filtering, and retrieval. *IEEE Trans. Multimedia 3*, 1 (2001), 141–151.

[29] Pentland, A., Picard, R., and Sclaroff, S. Photobook: Tools for content-based manipulation of image databases. In *SPIE, Storage and Retrieval of Image and Video Databases II* (1994), pp. 2185–05.

[30] Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 1994.

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

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

[33] Santini, S. Semantic modalities in content-based retrieval. In *IEEE International Conference on Multimedia and Expo* (New York City, USA, July 2000), vol. II.

[34] Sobchack, T. and Sobchack, V. *An Introduction to Film*. Scot, Foresman and Company, 1987.

[35] Srinivasan, M., Venkatesh, S., and Hosie, R. Qualitative extraction of camera parameters. *Pattern Recognition 30*, 4 (1997), 593–606.

[36] Truong, B.T., Dorai, C., and Venkatesh, S. New enhancements to cut, fade, and dissolve detection processes in video segmentation. In *Proceedings of the 8th ACM International Conference on Multimedia* (Los Angeles, California, November 2000), pp. 219–227.

[37] Vasconcelos, N. and Lippman, A. A Bayesian video modeling framework for shot segmentation and content characterization. In *CVPR'97* (San Juan, Puerto Rico, 1997).

[38] Vasconcelos, N. and Lippman, A. Bayesian modeling of video editing and structure: Semantic features for video summarization and browsing. In *ICIP'98* (Chicago, Illinois, 1998).

[39] Vendrig, J., Worring, M., and Smeulders, A. Model based interactive story unit segmentation. In *IEEE International Conference on Multimedia and Expo 2001, ICME2001* (Tokyo, Japan, August 2001), pp. 1084–1087.

[40] Vogler, C. *The Writer's Journey: Mythic Structure for Storytellers and Screenwriters*. Michael Wiese Productions, 1992.

[41] Vogler, C. *The Writer's Journey: Mythic Structure for Storytellers and Screenwriters, Revised and Expanded*. Pan Books, 1999.

[42] Yoshitaka, A., Ishii, T., Hirakawa, M., and Ichikawa, T. Content-based retrieval of video data by the grammar of film. In *IEEE Symposium on Visual Languages* (Capri, Italy, 1997).

[43] Zettl, H. *Sight, Sound, Motion: Applied Media Aesthetics*. Wadsworth Pub Co., 1973.

[44] Zhao, L., Yang, S.-Q., and Feng, B. Video scene detection using slide windows method based on temporal constraint shot similarity. In *IEEE International Conference on Multimedia and Expo 2001, ICME2001* (Tokyo, Japan, August 2001), pp. 649–652.