Bridging the Semantic Gap in Content Management Systems: Computational Media Aesthetics

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

Svetha Venkatesh
Department of Computer Science
Curtin University of Technology
GPO Box U1987, Perth, 6845, W. Australia
svetha@cs.curtin.edu.au

ABSTRACT

With the explosion of online media and media-based services, a key challenge in the area of media management is automation of content annotation, indexing, and organization for efficient access, search, retrieval, and browsing applications. One of the major failings of current media annotation systems is the semantic gap which refers to the discontinuity between the simplicity of features or content descriptions that can be currently computed automatically and the richness of semantics in user queries posed for media search and retrieval. This paper proposes an approach that targets at bridging the semantic gap and building innovative content annotation and navigation services. The approach is founded upon an understanding of media elements and their role in synthesis and manipulation of program content with a systematic study of media productions. It proposes a framework for computational understanding of the dynamic nature of the narrative structure and techniques via analysis of the integration and sequencing of audio/visual elements. The resulting system will lead to automatic content organization and interpretation that provides high level and high quality content descriptions to aid in search, retrieval, and browsing and also to objective and consistent distillation of the common features of successful audio-visual strategies.

1. INTRODUCTION

While issues of media archival as well as delivery on the Internet and corporate intranets are adequately addressed by improved compression standards, faster networks, and advances made in storage and streaming technologies, the challenges of automating media annotation, content indexing, segmentation, and organization for search, retrieval, and browsing applications are still being tackled. Automatic content indexing and annotation is a growing area of research in media computing, and a recent survey paper summarizes the state of art and identifies the key challenges [14]. The failing of current systems is that while "the user seeks semantic similarity, the database can only provide similarity on data processing". The authors define the semantic gap as the "lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data has for a user in a given situation" [14]. The discontinuity between the simplicity of features or content descriptions that can be currently computed automatically and the richness of semantics in user queries posed for media search and retrieval makes user acceptance and adoption of automated content annotation systems very difficult. The authors of the survey conclude that "bridging the semantic gap between the simplicity of available visual features and the richness of user semantics" is the key issue in building effective content management systems.

To address this issue, we depart from existing approaches to deriving video content descriptions [8, 12, 11, 13]. Motivated and directed by video production principles, we propose an approach that goes beyond representing what is being directly shown in a video or a movie, and aims to understand the semantics of the content portrayed and to harness the emotional, visual appeal of the content seen. It focuses on deriving a computational scheme to analyze and understand the content of video and its *form*. Accepted rules and techniques in video production are used by directors worldwide to solve problems presented by the task of transforming a story from a written script to a captivating narration [5]. These rules, termed as *film grammar* in the movie domain, refer to repeated use of certain objects, visual imagery, and patterns in many films to instantly invoke a specific cinematic experience to the viewers [16, 15]. The rules and icons serve as shorthand for compressing story information, characters, and themes into known familiar formulae, often becoming the elements of a genre production. They constitute a style or form of artistic expression that is characteristic of content portrayed, and can be considered to be almost idiomatic in the language of any program composer or director. Production rules are found more in history of use, than in an abstract predefined set of regulations, and elucidate on ways in which basic visual and aural elements can be synthesized into larger structures.

We hypothesize that the employment of these tacitly followed rules in any genre not only can be understood and derived automatically with a systematic study of media productions, but also be exploited in characterizing what is happening in a video for high-level video/film abstraction in an algorithmic framework we term, *Computational Media Aesthetics*. The framework allows for a computational understanding of the dynamic nature of the narrative structure and techniques via analysis of the integration and sequencing of audio/visual elements, and is targeted at bridging the semantic gap and building effective content management systems at higher levels of abstraction and meaning. Further, it puts video/film anal-

ysis on a sound footing resting on principles and practices from video/film production rather than on ad hoc schemes. While earlier work [7, 9] using film grammar has focused on content generation, synthesis of video presentations, and virtual worlds, our emphasis is on characterizing, describing, and structuring of produced videos for media search, segment location and navigation services.

The layout of this paper is as follows: in Section 2 we define and discuss the framework, Computational Media Aesthetics. The lessons we learned in the use of this framework to extract two higher order semantic constructs from film are discussed in Section 3. Our conclusions follow in Section 4.

2. COMPUTATIONAL MEDIA AESTHETICS

Zettl defines Media Aesthetics as a study and analysis of media elements such as lighting, motion, colour and sound both by themselves and their roles in synthesizing effective productions [16]. We define *computational media aesthetics* as the algorithmic study of a number of image and aural elements in media and the computational analysis of the principles that have emerged underlying their use and manipulation, individually or jointly, in the creative art of clarifying, intensifying, and interpreting some event for the audience.

What does this new framework entail? By focusing on the emotional and visual appeal of the content, it attempts to uncover the semantic and semiotic information by a study of the relations between the cinematic elements and narrative form. It enables distilling techniques and criteria to create efficient, effective and predictable messages in media communications, and to provide a handle on interpreting and evaluating relative communication effectiveness of media elements through a knowledge of film codes that mediate perception, appreciation and rejection.

Our computational framework analyses and understands the content of video and its form. This approach, undergirded by the broad rules and conventions of content creation, uses the production knowledge to elucidate the relationships between the many ways in which basic visual and aural elements are manipulated in video and their intended meaning and perceived impact on content users. Our computational scheme analyzes videos to understand the film grammar, in particular and uses the set of rules that are commonly followed during the narration of a story, to assist us in deriving the annotation or description of video contents effectively. A system built using this principled approach where videos are analyzed guided by the tenets of film grammar will be effective in providing high-level concept oriented media descriptions that can function across many contexts and in enhancing the quality and richness of descriptions derived. We propose a two-tiered framework: Primitive feature extraction and a complex higher order semantic construct extraction stage (See Figure 1).

2.1 Primitive Feature Extraction

In our approach, first, like those of other researchers, simple features such as colour, motion, editing effects, sound signal energy, etc are extracted. Given a video of a movie, news program, a class or a training program, shot segmentation is carried out to partition the video into atomic units for further processing. Based on the low level visual and aural attributes, various shot-based attributes are computed: shot duration, average number of shots per unit time, its variance, shot colour features, average shot motion and variance, changes in perceived visual motion, shot audio energy level, etc. These features can be directly computed from frame or shot processing.

2.2 Higher Order Semantic Construct Extraction

What sets the framework apart from other schemes is this stage. Here, we extract complex constructs, or expressive elements that expose the underlying semantic information embedded in the media production. The extraction of increasingly complex features from a hierarchical integration of underlying primitives is a commonly followed approach. But the key difference is this framework of analysis based on production knowledge, that is, to both define what to extract, and how to extract these constructs we seek guidance from film grammar. We do so because directors create and manipulate expressive elements related to some aspect of visual or emotional appeal in particular ways to have maximum impact. With movies for example, we draw attention to the film creation process, and argue that to interpret the data one must see it through the filmmaker's eye. Film grammar is the portal that gives us insight into the film creation process. It can tell us not only what expressive elements a director manipulates, but also how she does it, why, and what the intended impact is. Thus, complex constructs are both defined and extracted only if media production knowledge tells us that it is an element that the director crafts or manipulates intentionally. These elements by their derivation and study result in crafting human-friendly content descriptions since they directly impact viewers' engagement with the content portrayed.

These complex constructs typically cannot be extracted directly from the shots, like the primitive features, but are built upon them. Many cinematic techniques, operating on many media elements such as color, camera movements, and sound contribute to the creation of an expressive element, and therefore an integrated analysis of multiple low level features across sequences of shots becomes essential. We are aware that video production grammar may indeed lead us to some expressive elements that do not easily translate into algorithms. However, several expressive elements are based upon manipulation of physical elements such as objects and cameras, and therefore can be defined in terms of their operations and extracted. Pace, rhythm, and tone are examples of some of the first set of complex constructs that can be examined and studied.

3. CLOSE ENCOUNTERS WITH EXPRESSIVE ELEMENTS IN FILM

To explore the feasibility of our framework and test its efficacy we have investigated the following expressive elements in film: pace and rhythm. We outline in each case what we learned from film grammar, and then present our observations related to the extraction of these two expressive elements.

3.1 Pace, According to Film Grammar

Tempo or Pace is often used interchangeably in film appreciation, and refers to the "rate of performance or delivery". Zettl [16] makes a distinction in defining pace as the perceived speed and tempo as the perceived duration. Thus tempo/pace is a reflection of both the speed and time of the underlying events being portrayed and affects the overall sense of time of a movie. Tempo is crafted and manipulated in different ways. One technique is the montage that allows a director to manipulate the shot lengths used in the creation of a scene, thus deliberately controlling the speed at which a viewer's

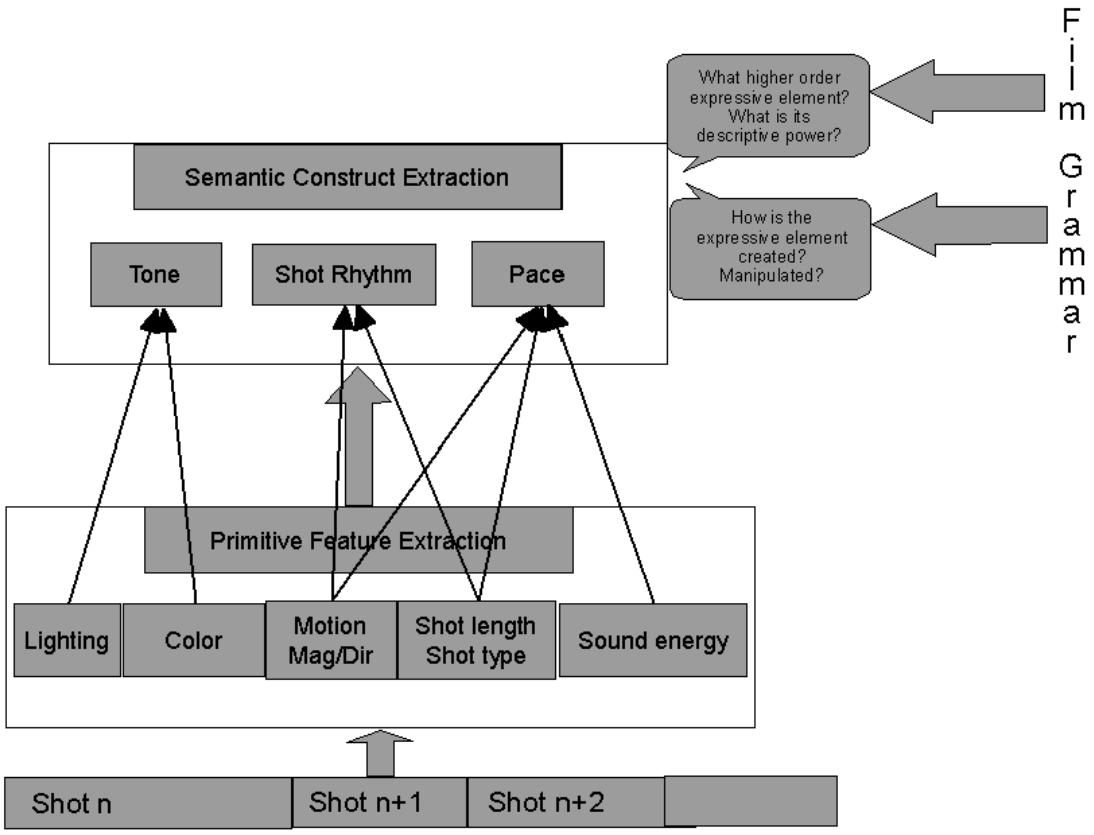

Figure 1: The Computational Media Aesthetics framework.

attention is directed. Another means by which a viewer's perception of speed can be manipulated is through controlling object and camera motion. Fast motion gives us the feeling of fast events, while no or little motion has the opposite effect on our perception of pace. Film audio is a third factor that increases or decreases our sense of the performance delivery. There may be other more subtle factors besides the story itself, but we argue that one can construct a computable and powerful expressive element, pace, that reasonably captures the flow of time in a movie based on the underlying primitives of shot length and motion.

3.2 Algorithms and Implications

Based on our understanding from film grammar that pace is primarily affected by shot length and pace, we define ***P(n)***, a continuous valued pace function as

$$\boldsymbol{P(n)} = \alpha(\, W(s(n))) + \frac{\beta(m(n) - \mu_m)}{\sigma_m}. \qquad (1)$$

where s refers to shot length in frames, m to motion magnitude, μ_m and σ_m, to the mean and standard deviation of motion respectively and n to shot number. The weights α and β, refer to relative weights that affect the extent to which shot length and motion contribute to pace. Without any *a priori* knowledge, they can be given values of 1, effectively assuming that both shot length and motion contribute equally to the perception of pace for a given film. It is possible, however, that under certain circumstances one or the other of these two impact more heavily on the audience perception of time, depending on the movie genre or a director's style. $W(s(n))$ is an overall two part shot length normalizing scheme, having the property of being more sensitive near the median shot length, but slows in gradient as shot length increases into the "longer" range.

We have examined this function for several motion pictures including the Titanic, Colour Purple, Lethal Weapon 2 and The Matrix, and can make the following observations:

- The ebb and flow of the pace function ***P(n)*** delineates the dramatic flow of the content and concurs very reliably with a qualitative assessment of movie tempo. The pace function paves way for characterizing movie content in many interesting ways, such as to organize content in terms of their dramatic import, to quantitatively compare and summarize

movie sections based on their tempo measure, or even to institute different policies to reduce our graduated tempo measure to labels, if desired to textually annotate content.

- Significant pace changes occur at the boundary of story sections, and often coincide with events of high dramatic import. Relative pace changes across the movie can be determined from the pace function. We extracted the edges of the pace function with a multi-resolution analysis, and found that large pace transitions coincide with what we term as story transitions. Smaller pace transitions coincide with what we term events, and these are differentiated from story transitions in that they occur more locally and contained in time. For example, the pace transition between the first and third class party scenes in the Titanic is a large one. The first class party scene is shown as sedate and slow. The third class party is vivacious and merry. As they happen juxtaposed, the pace difference between these two story sections is big and well captured. On the other hand, when Jack grabs Rose as she attempts to fall off the ship, there is a brief flurry of action, leading to a smaller pace transition which we label as a localized event. Our analysis has revealed that we can detect pace edges efficiently and that both story sections and events can be located (for further detail see [3, 2, 1]).

- Pace, therefore, is found to be high-level and fundamental (applicable to multiple contexts), yet manifest in a way to be computationally tractable. It offers pointers to automatically organizing videos into higher-level segments characterized by their dramatic development and to their high-level descriptions.

Results from The Matrix, one of a number of movies analyzed are presented here for the purpose of demonstration. Figure 2 shows the pace plot of a section of the movie with located edges indicated for two of the 4 Σ/τ combinations used in edge detection, and Table 1 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 1: Labelled story sections and events identified from tempo changes in The Matrix (cf. Figure 2).

Story Element detected (high thresh)	
A	Morphius in captivity
B	The rescue of Morhpius
C	Neo finally confronts the matrix
D	Neo is the 'one'
Event detected (low thresh)	
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

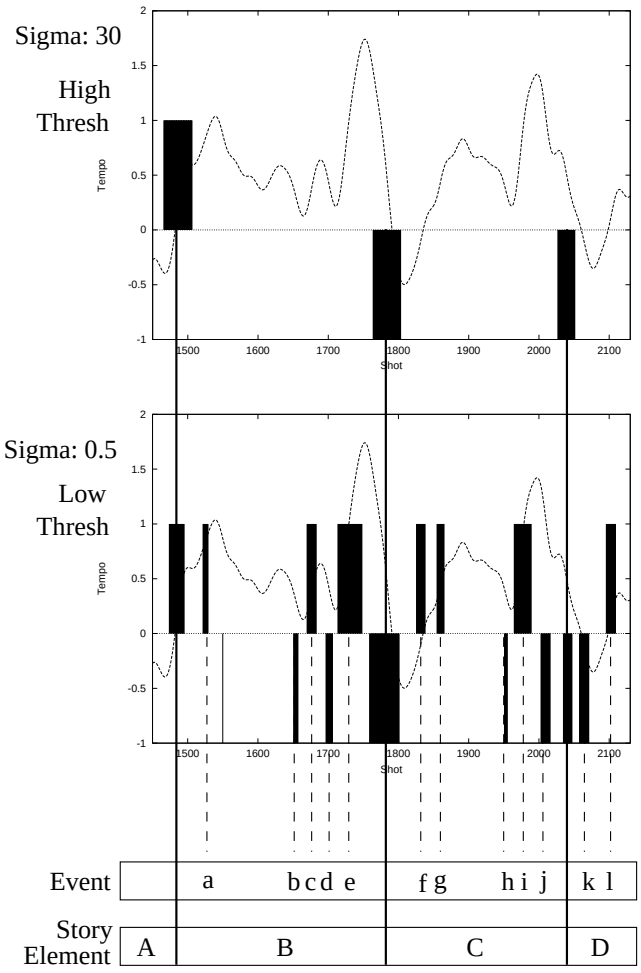

Figure 2: Results of edge detection on pace flow and corresponding story sections and events from the Matrix.

3.3 Rhythm in Film

Film rhythm is another complex narrative concept used to endow structure and form to film. Mitry defines it as an "organization of time" [10]. Of the many, often elusive cinematic devices contributing to film rhythm, Bordwell and Thompson [6] state that "frame mobility involves time as well as space, and film makers have realized that our sense of duration and rhythm is affected by the mobile frame". They list camera position/movement, sound rhythm, and editing as constituent elements of rhythm. Further they label resulting rhythms types in higher level terms by stating that a "camera motion can be fluid, staccato, hesitant and so on".

Thus, because a film is structured in time with editing, it manifests a natural beat, and has an intrinsic rhythm. To find this rhythm, one must examine a neighborhood of shots. In addition, since both shot length and motion contribute to rhythm, one can examine the rhythm that arises individually and jointly from these contributing elements. Since shot length and motion are computable, motion rhythm and editing rhythm are likewise derivable from them.

3.4 Computing Rhythm

We have chosen to study two main contributing factors to rhythm, namely the shot length and motion. We examine each separately (see [4] for details).

Shot rhythm: Since shot lengths can be crafted in many ways to produce almost an infinite variety of beats, we restricted the initial shot rhythm classes to the following:

- Metrical: in which a series of shots have nearly the same shot length.
- Attack: in which a series of shots have steadily decreasing shot length.
- Decay: corresponds to monotonically increasing shot length.
- Free: None of the above, and therefore in these sections of the movie, shot rhythm is secondary instead of being the driving factor behind shot placement and length.

We implemented an algorithm to extract and classify shot neighbourhoods with their shot rhythm label. We have experimented with both the neighbourhood size and the tolerances required to classify movie sections robustly. We make the following observations: (a) Attack is often used to lead into a dramatic event. A rising shot rate captures and leads the viewer to the events of high dramatic import that are about to follow. (b) It does not appear that decay is crafted to the same extent as attack, but it has the opposite effect to attack in perception. (c) Even if the identification of attack, metric and decay sections does not tell us about the exact content, they act as clear signs that the director is creating or reinforcing *something* by their use. (d) A precision of about 70-75% was registered in automatically classifying sections with these shot rhythm labels.

Motion Rhythm: Film grammar guides us to divide the classes of motion rhythm as

- No Motion,
- Fluid: Where the motion is smoothly changing or constant, and
- Staccato: where the motion contains abrupt speed changes.

A shot motion behaviour classifier was implemented and the results analyzed across a range of movies. Our observations from these experiments are: (a) The classifier in labelling the motion rhythm classes achieved an accuracy of 70-80%. (b) The most interesting results from examining motion rhythms arose from the degree of common scene content found in different rhythm categories. (c) Neighbourhoods with predominantly fluid rhythm generally correspond to long and establishing shots. (d) Staccato motion neighbourhoods are very taxing and used sparingly, particularly in extremely violent scenes. (e) A mixture of Fluid and Staccato correspond to scenes where a dual perspective is being shown, for example, action sequences.

3.5 Inferring Content Semantics

What inferences can we draw by a joint analysis of both shot and motion rhythms? How do they interact? What clues do they provide for automatically structuring content? To answer these questions, we studied both shot and motion rhythms across several movies. Our main conclusions are the following:

It is not possible to make a deduction on the actual content of a scene from its rhythm. However, one can reliably infer that when rhythm changes, something has changed in the scene.

We can also conclude something about why the content has changed. A shot device (either metric, attack or decay rhythm) accompanied by a motion rhythm change generally corresponds to a scene change being precipitated by an event, or to the reinforcement of change. A free shot rhythm accompanied by a motion rhythm change generally implies a change to a different locale, scene or sequence abruptly. A shot device (metric, attack or decay) with no accompanied motion change generally implies the same film setting or a result of an event.

Rhythm, too therefore, like pace, is computable. It leads us to understand where scene content has changed, and hypothesize as to why this could have occurred. This change and its reasoning is at a much higher semantic level than could have been done reliably by merely examining the change of a low level primitive like shot length.

These experiments present some of our initial findings. We are currently gathering experimental data to support distinctive occurrences of shot and motion rhythm classes across genres, in order to understand whether we can index the content at a finer resolution.

4. CONCLUSIONS

We propose to bridge the semantic gap by using media production understanding to guide media analysis. We believe that it is this understanding that will enable us to both formulate and extract the correct semantic entities, and enable us to structure video and film automatically. These semantics can lead to the development of shared vocabularies for structuring video and film, and serve as the foundation for media description interfaces. While we have drawn upon film grammar to derive the expressive elements in film, we believe that such an approach will work in other video domains. News, Sitcoms, Sport etc. all have more or less complex grammars that may be used to capture their crafted structure. There is structure regardless of particular media context but there may not be homogeneity, and therefore it helps to be guided by production

knowledge in media computing.

We are building upon our initial results to eventually develop video-watching agents or systems that can make human-like judgements about images and sounds in film, and will lead to finely tuned, personalized screening and practical filtering tools. The software models will enable technologies that can emulate human perceptual capabilities on a host of difficult tasks such as parsing movie into sections of interest, making inferences about movie semantics and about the perceptual effectiveness of the messages contained, and estimating the similarity of two productions.

5. ACKNOWLEDGEMENTS

We thank Brett Adams for being a willing and helpful partner in our close encounters with the expressive elements explored so far.

6. REFERENCES

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

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

[3] B. Adams, C. Dorai, and S. Venkatesh. Towards automatic extraction of expressive elements from motion pictures: Tempo. In *IEEE International Conference on Multimedia and Expo*, volume II, pages 641–645, New York City, USA, July 2000.

[4] B. Adams, C. Dorai, and S. Venkatesh. Automated film rhythm extraction for scene analysis. In *IEEE International Conference on Multimedia and Expo*, Tokyo, Japan, August 2001, To appear.

[5] D. Arijon. *Grammar of the film language*. Silman-James Press, 1976.

[6] D. Bordwell and K. Thompson. *Film Art, 5th Ed.* McGraw-Hill, 1997.

[7] M. Davis. Knowledge representation for video. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 120–127, 1994.

[8] J. Huang, Z. Liu, and Y. Wang. Integration of audio and visual information for contend-based video segmentation. In *Proc. IEEE International Conference on Image Processing*. IEEE, 1998.

[9] C. Lindley. A computational semiotic framework for interactive cinematic virtual worlds. In *Workshop on Computational Semiotics for New Media*, Guildford, Surrey, UK, 2000.

[10] J. Mitry. *The Aesthetics and Psychology of the Cinema*. The Athlone Press, London, 1998.

[11] J. Nam, M. Alghoniemy, , and A. Tewfik. Audio visual content based violent scene characterization. In *Proc. IEEE International Conference on Image Processing*. IEEE, 1998.

[12] V. Pavlovic. Multimodal tracking and classification of audio visual features. In *Proc. IEEE International Conference on Image Processing*. IEEE, 1998.

[13] C. Saraceno and R. Leonardi. Identification of story units in audio visual sequences by joint audio and video processing. In *International Conference on Image Processing*. IEEE, 1998.

[14] A. Smeulders, M. Worring, S. Santini, and A. Gupta. Content based image retrieval at the end of the early years. *pami*, 22(12):1349–1380, 2000.

[15] T. Sobchack and V. Sobchack. *An introduction to film*. Scot, Foresman and Company, 1987.

[16] H. Zettl. *Sight, Sound, Motion: Applied Media Aesthetics*. 3rd Edition, Wadsworth Pub Company, 1999.