RC22656 (W0211-278) November 27, 2002
Computer Science

IBM Research Report

Bridging the Semantic Gap in Content Management Systems

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

Svetha Venkatesh
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 1

BRIDGING THE SEMANTIC GAP IN CONTENT MANAGEMENT SYSTEMS

Computational Media Aesthetics

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, Western Australia
svetha@cs.curtin.edu.au

Abstract

With the explosion of digital media and online services, a key challenge in the area of media management is automation of content annotation, indexing, and organization for efficient media access, search, retrieval, and browsing. A major failing of current media annotation systems is the semantic gap — the incompatibility between the low-level features that can be currently computed automatically to describe media content and the high-level meaning associated with the content by users in media search and retrieval. This inevitably leads to the problem of content management systems returning media clips that are similar to one another in terms of low-level descriptions, but are completely different in terms of semantics sought by the users in their search. This chapter introduces Computational Media Aesthetics as an approach to bridging the semantic gap, outlines its foundations in media production principles, presents a computational framework to deriving high-level semantic constructs from media, and describes the structure of this collection.

Keywords:

Media archives, digital content management, video indexing, content-based search and annotation, semantic indexing, MPEG-7, Computational Media Aesthetics, semantic gap, production knowledge, film grammar

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 the art and identifies the key challenges [5]. 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" [5]. The discontinuity between the low level features or content descriptions that can be currently computed automatically and the rich 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.

Motivated and directed by media 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 [1]. These rules, termed as *film grammar* in the movie domain, refer to repeated use of certain objects, visual imagery, and practices and patterns in many films to instantly invoke a specific cinematic experience to the viewers [7, 6]. The rules and icons serve as shorthand for compressing story information, characters, and themes into known familiar formulae, often becoming the common 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. Movie 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 advocate 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 media analysis on a sound footing resting on principles and practices from media production rather than on ad hoc schemes. While earlier work [2, 4] 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.

1. Computational Media Aesthetics

Zettl defines media aesthetics as a study and analysis of media elements such as lighting, motion, color and sound both by themselves and their roles in synthesizing effective productions [7]. 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 [3].

What does this new framework entail? By focusing on the emotional and audio-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.

This computational framework analyses and understands the content of video and its form. The approach, undergirded by the broad rules and conventions of content creation, uses media 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. It analyzes videos to understand the production grammar in particular and uses the set of rules that are followed during the narration of a story, to assist in deriving the high level description of video content 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 1.1).

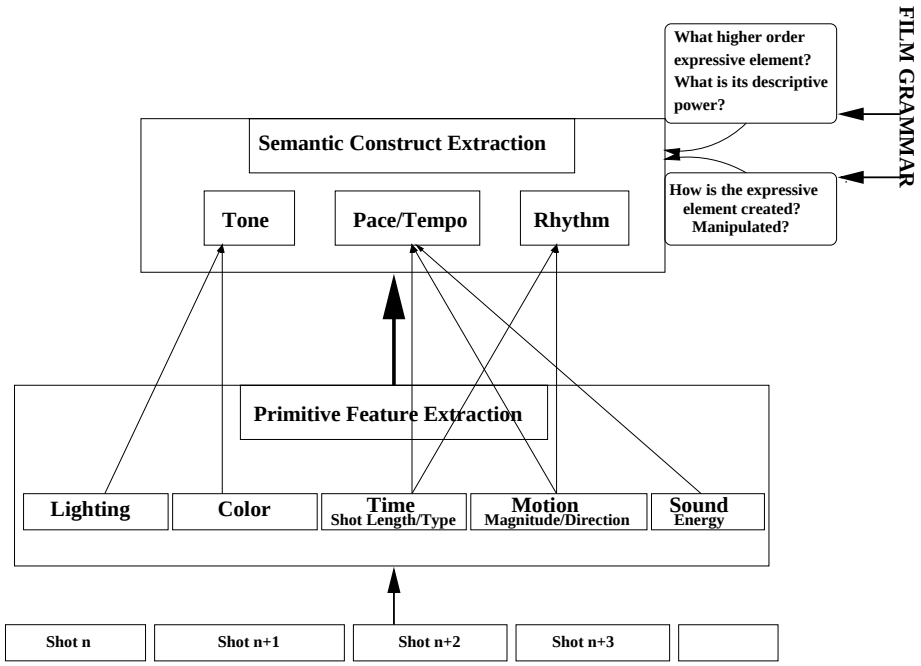

Figure 1.1. The Computational Media Aesthetics framework.

1.1 Primitive Feature Extraction

In our approach, first, like those of other researchers, simple features from various audio-visual elements such as color, motion, editing effects, sound signal, etc are extracted. Given a video of a movie, news, sports, or an education and training program, shot segmentation is carried out to partition the video into basic units of shots for further processing. Based on visual and audio devices, various low level shot-based attributes are computed: Shot duration, average number of shots per unit time, its variance, shot color features, average

shot motion and variance, changes in perceived visual motion, shot audio energy level, etc. These features can be directly computed from frames or shots.

1.2 Higher Order Semantic Construct Extraction

What sets this framework apart from other schemes is this stage. Here, we extract complex constructs or expressive elements that expose the underlying semantic information embedded in media productions. The extraction of increasingly complex features from a hierarchical integration of underlying primitives is a commonly followed approach for video analysis. 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 and logic 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 film 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 manifests and manipulates, but also how she does it, why, and what the intended meaning and impact is. Thus, complex constructs are both defined and extracted only if media production knowledge tells us that they are elements that the director crafts or manipulates intentionally. These semantic constructs by their derivation and examination lead to deriving human-friendly content descriptions since they directly impact viewers' engagement with the content portrayed.

What we are advocating is a return to the *roots of production*. To endow high-level semantics to the annotations constructed, we are pursuing an integrated analysis of multiple low level features firmly guided by the production techniques and logic used by the creators of film. The essence of Computational Media Aesthetics is that in order to create tools and technologies for automatically understanding and annotating media, we need to be able to interpret the data with its maker's eye, that is, to faithfully reflect the forces at play in media construction.

2. What is This Book About?

The spirit of this book is captured in the words of Sieng Tisang:

To return to the root,
Is to find meaning.
To pursue appearances,
Is to miss the source.

In the first chapter, we describe the problem of semantic gap in media management systems, and present Computational Media Aesthetics as a principled and sound approach to address this fundamental problem. We describe the voyage to examine the roots of media production and aesthetic principles of creating content in order to establish the foundation for semantic analysis of media and to formulate new ways for automatic media processing.

Zettl starts us on this journey by taking us right back to the source of production and film grammar and details the role of media elements in shaping a message in film and television. He lays the road-map for this book by introducing us to the basic aesthetic elements: Light and color, two-dimensional space, three-dimensional space, time-motion and sound. Focusing on the basic properties and their structural potential, he elaborates on the role of these fundamental aesthetic elements and the common conventions used to shape meaning. The rest of the book takes us further in this voyage, anchoring us with studies of different aesthetic elements along the way.

Mamber inquires into the roots of production deeply when he asks us to look at "the work itself as a browsing tool". Using the Cartesian grid as an anchor, he presents his vision for building frameworks for media exploration and navigation.

The next two chapters explore the workings of Computational Media Aesthetics when applied to extraction of meaning using two of the aesthetic elements introduced by Zettl: Time and color. Adams et al. explore film tempo, and show us that though descriptive and sometimes fuzzy in scope, film grammar gives us rich insights into the perception of subjective time and its manipulation by the makers of film for drama. Building on this, they define film tempo as a computable entity, and explores the extraction and use of tempo in delineating dramatic story sections and events in motion pictures.

Del Bimbo et al. embark on an understanding of the dynamics of color in video and examine its potential as an expressive element. Using commercials as the domain, they examine how the flow of color over time creates different signatures that are unique for different types of commercials. They formulate robust techniques for the extraction of color dynamics and demonstrate the existence of these signatures with experiments on different kinds of commercials.

The following two chapters open up a portal to sound, exposing the intricacies of manipulating sound as a structural element in media. Focusing on the production viewpoint of constructing scenes with audio, Pfeiffer and Srinivasan elucidate current audio editing practices and the cinematic techniques that are used to create scene structures in motion pictures. They elaborate on current computational approaches to scene determination using audio features and sound cues. Following this survey of the state of the art, Moncrieff et al. explore a specific audio element: Sound energy dynamics and its manipulation to effect drama in film. Using horror films as a domain of study, they outline

techniques for the extraction of dynamical features from sound energy flow and examine their role in influencing viewer emotions such as surprise and alarm.

The book concludes with Nack's vision for the future for media computing. He argues that a framework that merges media generation and annotation is the future of what he terms "media aware knowledge spaces". Using the domains of digital production and encyclopaedic spaces, he gives insights into the development of tools and outlines their use in facilitating media search, navigation and discourse between communities of varying expertise.

This book gives us exciting glimpses into this nascent, interdisciplinary field of research. We are aware that video production grammar may indeed lead us to some expressive elements that do not easily translate into computation and algorithms. The production knowledge is very much descriptive rather than prescriptive in nature. As such it should be used with a degree of caution and flexibility. There is a strong recognition, however, that filmmakers generally ignore the tenets of film grammar at the risk of confusing their audience, and thus reduce the ability of their film to communicate their desired intent (unless of course, the filmmaker is seeking to communicate confusion). Therefore, there is much to be gained from the realities of the media creation process and the way of Computational Media Aesthetics shows great promise. If the results so far seem limited, it is because of the inherent complexity of the area, best described in the words of Blaise Pascal,

" And everything is both cause and effect, working and worked upon, mediate and immediate, all things mutually dependent."

The aspiration for the future then is to extract the quintessence of the digital medium.

References

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

[2] Davis, M. Knowledge representation for video. In *Proceedings of the National Conference on Artificial Intelligence* (1994), vol. 1, pp. 120–127.

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

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

[5] Smeulders, A., Worring, M., Santini, S., and Gupta, A. Content based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 12 (2000), pp. 1349–1380.

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

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