

Computational Media Aesthetics: Finding Meaning Beautiful

Chitra Dorai
IBM T.J. Watson
Research Center

Svetha Venkatesh
Curtin University
of Technology

Content management's future is bright. Innovative media management, annotation, delivery, and navigation services will enrich online shopping, help-desk services, and anytime-anywhere training over wireless devices. Semantics-based annotations will break the traditional linear manner of accessing and browsing media and will support vignette-oriented access of audio and video. This can lead to new offerings of customized media-management utilities for various market segments such as online education and training, advertising, news networks, and broadcasting studios.

However, the semantic gap between the rich meaning that users want when they query and browse media and the shallowness of the content descriptions that we can actually compute is weakening today's automatic content-annotation systems. This is a crucial obstacle that we must overcome to achieve that bright future. A serious need exists to develop algorithms and technologies that can annotate content with deep semantics and establish semantic connections between media's form and function, for the first time letting users access indexed media and navigate content in unforeseeable and surprising ways.

To address these underlying problems, we advocate an approach that markedly departs from existing methods¹ based on detecting and annotating low-level audio-visual features. To go beyond representing what a video or movie directly shows, we postulate that we must analyze and interpret the content's visual, aural, and emotional impact. Our contention is that we must understand compositional and aesthetic media principles to guide content analysis.

The inspiration

What avenues do we have for analyzing and interpreting media? Structuralism,² in film studies for example, proposes film segmentation followed by an analysis of parts or sections. We can see a

structuralistic approach to media computing in the recent emphasis on the underlying "meaning of the established relations between the single components of a multimedia system and exposing the main semantic and semiotic information hidden in the system's unified structure."³ We can treat structural elements, or portions of a video, when divested of cultural and social connotations, as plain data and therefore examine them using statistical and algorithmic tools for content description.

A richer and more important source is production knowledge or film grammar. Directors worldwide use accepted rules and techniques to solve problems in transforming a story from a written script to a captivating visual and aural narration.⁴ These rules encompass a spectrum of cinematic aspects ranging from shot arrangements, editing patterns, and the triangular camera placement principle to norms for camera motion and action scenes. Over time, codes and conventions used to narrate a story with a certain series of images have become so standardized and pervasive that they now appear natural to modern film viewers.

However, video production mores are found more in history than in an abstract predefined set of regulations; they're descriptive rather than prescriptive. They elucidate ways in which we can synthesize basic visual and aural elements into larger structures and the relationships that exist between the many cinematic techniques employed worldwide, their intended meaning, and their emotional impact on movie audiences.

The way

Media aesthetics is a process of examining media elements such as lighting, picture composition, and sound—by themselves or jointly—and a study of their roles in manipulating our perceptual reactions, communicating messages artistically, and synthesizing effective media productions.⁵ Inspired by that process, we defined *computational media aesthetics* as

the algorithmic study of a variety of image and aural elements in media (based on their use in film grammar). It is also the computational analysis of the principles that have emerged underlying their manipulation in the creative art of clarifying, intensifying, and interpreting an event for an audience.

Computational media aesthetics lets us distill successful techniques and criteria to create efficient, effective, and predictable messages in media communications. It provides a handle on interpreting and evaluating relative communication effectiveness of media elements with a knowledge of film codes that mediate perception, appreciation, and sometimes rejection. While the affective computing area aims to understand and enable computers to interpret and respond to users' emotional states, our approach aims to understand how directors use visual and sound elements to heighten the audience's emotional experience. Computational media aesthetics exposes the semantic and semiotic information embedded in media productions by focusing on the representation of perceived content in digital video and 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.

The importance of our approach is that it computationally analyzes the usage patterns of visual and aural elements in TV and film, which lets us create tools that facilitate efficient, effective, and predictable transformation of ideas into messages viewers can perceive. Understanding the dynamic nature of the narrative structure and techniques by analyzing the sequencing and integration of audio-visual elements will lead to better characterization of content and form. It will also let us develop tools for mass adoption of the successful techniques used in film and video making. In turn, this understanding will lead to high-level semantic annotation of movies and videos, and to a more effective user-query interpretation capability, allowing easier, human-oriented content description in queries submitted to media search engines. This new research area examines and associates deep semantics to the narrative structure in movies and television.

The difference

While other researchers have sought to model specific events in a particular video domain in detail, our research delineates the expressiveness of the content's visual and aural patterns and the thematic units highlighted by them (such as fast-

To create tools for automatically understanding video, we need to be able to interpret the data with its maker's eye.

paced, tranquil, or horror scenes) regardless of a story's specific nature. The essential difference lies in developing analytical techniques founded on production knowledge for film and video understanding. We hope to extract high-level semantics associated with the cinematic elements and narrative forms synthesized using them and illustrate how we can detect and reconstruct such high-level mappings by using software models.

Other researchers have used film grammar as a compositional framework in research related to content generation, synthesis of video presentations, and virtual worlds.^{6,7} They've also suggested an integrated framework for new media productions in which we can author and generate particular presentations in a variety of forms using a range of tools interacting with a distributed media repository.³ Our research systematically uses film grammar to inspire and underpin an automated process of analyzing, characterizing, and structuring professionally produced videos.

The first steps

We've created a framework for computationally determining elements of form and narrative structure in movies from the basic devices of film grammar—namely, the shot, motion, recording distances, and practices that are commonly followed during the structuring of a story's audio-visual narration. We first extract the content's primitive computable aspects. We then define and construct new expressive elements (higher order semantic entities) from the primitive features. We base the definition and extraction of these semantic entities on production knowledge and film grammar. We formulate these entities only if directors design them and manipulate them for increased emotional engagement with viewers. The primitive features and the higher order semantic notions form the vocabulary for automated film-content description.

As an example, consider our software model of

an expressive element in movies: tempo, or pace. (Strictly speaking, pace refers to perceived speed, and tempo refers to perceived duration.⁵ Because our work derives both, we use them together.) Tempo, the rate of performance or delivery, can be fundamental in a movie (and therefore widely applicable) and yet manifest itself in such a way as to be computationally tractable. Several factors can create and affect tempo, such as shot length, motion, zoom, sound, and of course the story itself. Based on two attributes, the shot length and motion, we define a continuous tempo function⁸ that quantifies the notion of subjective time in a film. With our software model, we define and derive tempo plots for full-length movies. We determine tempo changes with an edge-detection technique, leading to the automatic extraction of dramatic story sections and events signaled by their unique tempo. Our experimental results confirm the reliable detection of actual tempo transitions, which serve as high-level indices into a story's dramatic ebb and flow and narration in motion pictures.

Research in computational media aesthetics will entail the further fleshing out of our understanding and application of film grammar to media analysis. This will yield new aspects we must compute, improvements to existing measures, and insights into how we'll construct new tools from the knowledge gained.

The challenges

To create tools for automatically understanding video, we need to be able to interpret the data with its maker's eye. As a result, a number of challenging questions arise:

- Can we dynamically detect successful aesthetic principles with accuracy and consistency using computational analysis?
- Can we build new postproduction tools based on this analysis for rapid, cost-efficient, and effective moviemaking and consistent evaluation?
- How can we use these successful audio-visual strategies for improved training and education in mass communication?
- How do we raise the quality of media annotation and improve the usability of content-based video search and retrieval systems?

The future

Film isn't the only domain with a grammar we

can exploit. News, sitcoms, and sports all have more or less complex grammars that we can use to capture their crafted structure. In this new world of self-expression, there will soon be a desire to manipulate digital aesthetic elements to deliver messages in many different ways and a need to reverse engineer intent and meaning from available content. Computational media aesthetics takes us toward this goal. It lets us learn from the practitioners of artistic expression, to build tools and technologies for nonlinear media access and manipulation. We have a long way to go before we can design and deliver new forms of interactive media, but the future is bright. **MM**

Acknowledgments

We thank Brett Adams for his help in shaping our ideas and realizing them in concrete algorithms and a system implementation.

References

1. A. Smeulders et al., "Content-Based Image Retrieval at the End of the Early Years," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, Dec. 2000, pp. 1349-1380.
2. A.G. Antohin, *Semiotics 101*, Part II, Web document, 1999, <http://afronord.tripod.com/semio.html>.
3. F. Nack, "All Content Counts: The Future of Digital Media is Meta," *IEEE MultiMedia*, vol. 7, no. 3, July-Sept. 2000, pp. 10-13.
4. D. Arijon, *Grammar of the Film Language*, Silman-James Press, Los Angeles, Calif., 1976.
5. H. Zettl, *Sight Sound Motion*, third ed., Wadsworth Publishing, Belmont, Calif., 1999.
6. M. Davis, "Knowledge Representation for Video," *Proc. 12th Conf. Artificial Intelligence*, AAAI Press, Menlo Park, Calif., 1994, pp. 120-127.
7. C. Lindley, "A Computational Semiotic Framework for Interactive Cinematic Virtual Worlds," *Proc. Workshop Computational Semiotics for New Media*, Univ. of Surrey, UK, 2000.
8. B. Adams, C. Dorai, and S. Venkatesh, "Role of Shot Length in Characterising Tempo and Dramatic Story Sections in Motion Pictures," *Proc. IEEE Pacific Rim Conf. Multimedia*, IEEE Press, Piscataway, N.J., 2000, pp. 54-57.

Readers may contact Chitra Dorai at the IBM T.J. Watson Research Center, PO Box 704, Yorktown Heights, New York, NY 10598, email dorai@watson.ibm.com.

Contact Media Impact editor Frank Nack at CWI, Kruislaan 413, PO Box 94079, 1090 GB Amsterdam, The Netherlands, email Frank.Nack@cwi.nl.