M A R C D A V I S

School of Information Management and Systems
University of California at Berkeley

marc@sims.berkeley.edu
www.sims.berkeley.edu/~marc

P U B L I C A T I O N S

IDIC: Assembling Video Sequences from Story Plans and Content Annotations

Bibliographic Reference:

Marc Davis and Warren Sack. "IDIC: Assembling Video Sequences from Story Plans and Content Annotations." In: *Proceedings of IEEE International Conference on Multimedia Computing and Systems in Boston, Massachusetts*, IEEE Computer Society Press, 30-36, 1994.

IDIC: Assembling Video Sequences from Story Plans and Content Annotations

Warren Sack * and Marc Davis +

* MIT Media Lab, Machine Understanding Group, 20 Ames Street, Cambridge, MA 02139
phone: 617/253-9497 *email*: wsack@media.mit.edu

+ MIT Media Lab & Interval Research Corp., 1801 Page Mill Road, Building C, Palo Alto, CA 94304
phone: 415/354-3631 *email*: davis@interval.com

Abstract§

We describe a system, IDIC, which can generate a video sequence according to a story plan by selecting appropriate segments from an archive of annotated video. IDIC uses a simple planner to generate its stories. By critically examining the strengths and weaknesses of the representation and algorithm employed in the planner, we are able to describe some interesting similarities and differences between planning and video story generation. We use our analysis of IDIC to investigate the representation and processing issues involved in the development of video generation systems.

1. Introduction: The Common Sense of Television

Americans watch a lot of television. On average most watch six hours of TV a day, and most households have the set on for at least eight hours [Cross 1983, p. 2]. What are we learning from the attention we spend on soap operas, sitcoms, ads, Monday night football, talk shows, and music videos? A culturally specific form of common sense. Indeed what we are learning through the television has become, to a large extent, the consensual reality of the United States. Rodney King's beating by the L.A. police, the explosion of the space shuttle *Challenger*, former Vice-President Quayle's comments about *Murphy Brown*, and *Murphy Brown*'s response to Quayle, the name of *Lucy*'s husband (Ricky), and the slogan from the *Wendy*'s restaurant commercial which was often quoted in the 1984 presidential race ("Where's the beef?") are all examples of events which were seen by most of us, not with the naked eye, but on television; all of these events are "common sensical" to the extent that they are referents with which "everyone" is assumed to be familiar for the purposes of casual discourse. Ever since, at least, McCarthy's description of an *advice taker* [McCarthy 1958], a machine that could be programmed in a common vernacular, researchers (e.g., Lenat and Guha 1990; Hobbs and Moore 1985) have been trying to find a way to articulate "common sense" in a computationally interpretable form. It is striking that none of this research has been aimed at representing television, the subject which occupies almost as many of Americans' waking hours as work and school. One of our current concerns is to address this oversight. This paper is a description of some of our initial efforts aimed at articulating the "common sense" of television.

With our long-term research agenda we seek to address two issues: one technological and one theoretical:

- *The Technological Issue: Interactive Television:* In the next few years the technology of television will be integrated with computers. As a consequence, television (and also the "common sense" of television) will change. Viewers will have access to services which will allow them to search for and download movies and all types of television shows from distant sources. It will also be possible, with the advent of digital television, to program "interactive" shows which will allow the viewer to, for example, specify a change in narrative, replace characters or actors, specify camera movements, or, in general, to play the role, in a limited manner, of the director. In our research we are attempting to find the means to represent, index, and automatically draw inferences about television shows. We hope that this work will provide the underpinnings necessary to support the functionality of an interactive television technology.

§ Published in the *Proceedings of the IEEE International Conference on Multimedia Computing and Systems*, May 14-19, 1994, Boston, MA.

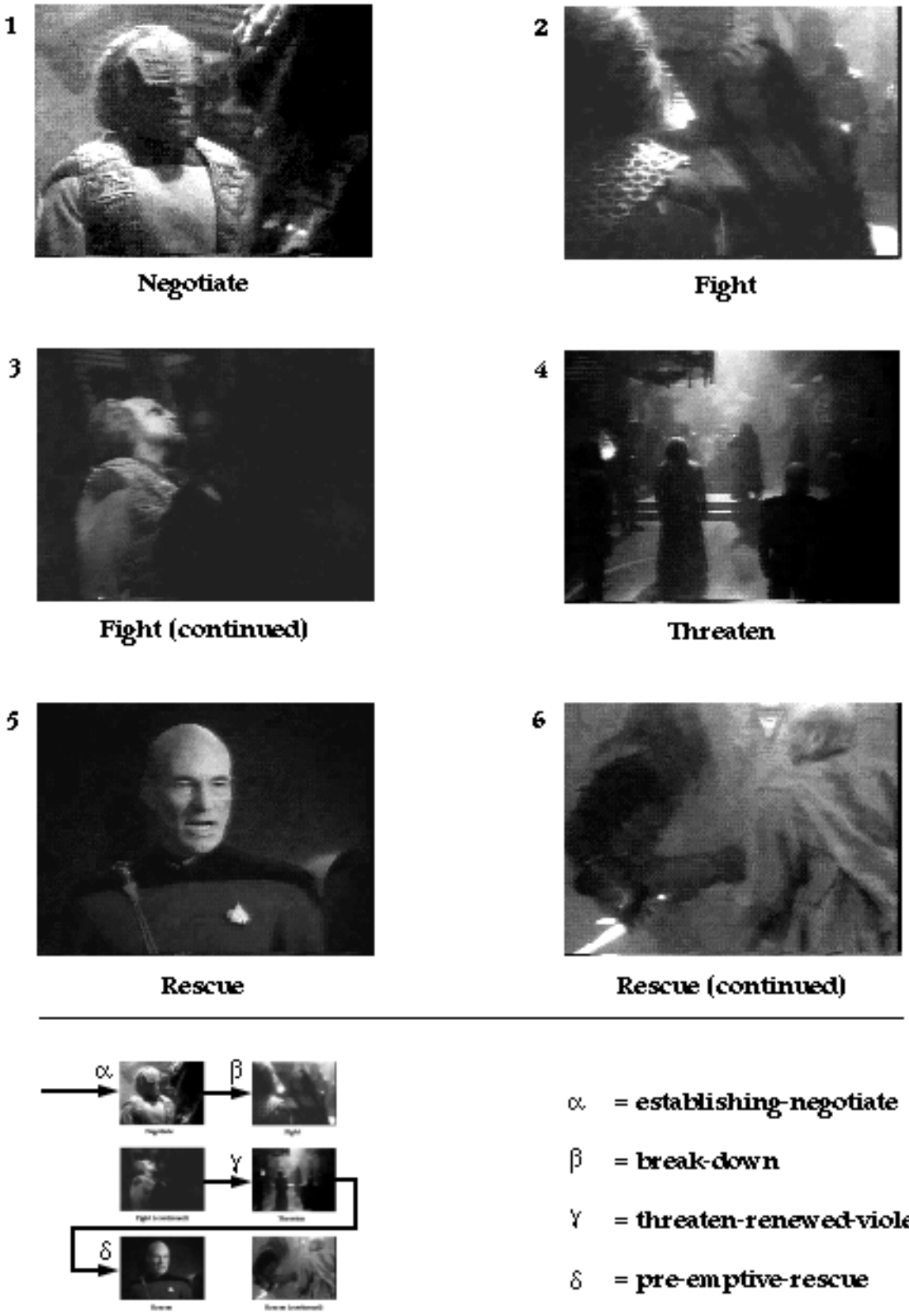

Figure 1: The "Rescue" Trailer

• *The Theoretical Issue: Television and AI Theories of Common Sense:* Within the discipline of artificial intelligence (AI) we often speak as though knowledge comes in only two flavors: (1) expert knowledge; and, (2) culturally independent "common sense" knowledge. Everyone is assumed to possess, at least some, "common sense." Thus, human novices, students, readers, viewers, or learners, in general, are prefigured, in the literature of AI as "non-experts;" i.e., as minds which possess the ubiquitous "common sense," but which lack a specific sort of knowledge, an expertise of a particular professional or academic discipline. This is an inadequate representation of "common sense" because it leaves no room for a study of the sorts of culturally specific and rarely archived knowledges that many of us are fluent in; e.g., popular culture. Consequently, we would contend that contemporary AI theories of representation are inadequate to the task of representing the "common sense" of television. The "common sense" of television is the content of television and the sort of learning and transformations experienced by viewers of television. In short, in AI it is difficult to construct flexible and perceptive representations of popular culture, in general, and television, in particular, because there exist no adequate means to represent the fact that producers and viewers know a lot of things which are neither as culturally independent as "common sense" has been presumed to be by AI researchers, nor as professionally or academically specialized as expert knowledge.

Our initial steps toward our long-term research goals have been, what we tend to refer to as, a "literature review by critical re-implementation." We are trying to reassess and extend older work in artificial intelligence (AI) to see if it is arguably applicable to the relatively unexamined domain of television. Our methodology involves re-implementing cognitive models as computer programs and then integrating them into larger systems for annotating, analyzing, and generating video. Instead of "writing off" older work, we are attempting to give ourselves first-hand experience with computer-based instantiations of prior research. Our aim has been to find a set of indexing and inferencing techniques which will allow us to create programs which can automatically create new videos by composing together parts of others stored in a digital archive. The work reported in the present paper was originally initiated to illustrate how planning techniques, as they have been described in the artificial intelligence literature, are *not* applicable to the task of video generation. Contradictorily, to our own surprise, we found that some planning techniques are indeed of interest in the domain of video generation.

This paper is divided into two sections.

(1) *An Example:* We give an example of the sort of videos that our simplest system can generate. This simplest of systems is nothing fancy: its inferencing capabilities are built upon a GPS-type [Newell and others 1963] planner. But, the system's output is of interest because it allows us to illustrate the sorts of mechanisms inherent to the domain of automatic video generation.

(2) *GPS and Video Generation:* We describe the architecture of our simplest system to point out the sources of the strengths and weaknesses illustrated by its output. Many arguments have been made in the AI literature to demonstrate that it is unrealistic to imagine that simple planning routines could ever do anything practical [Chapman 1987]. However, the analysis we provide of our system investigates how planning can be a tool for framing the problems of video generation: we find certain aspects of the representations used in planners (e.g., operators with add and delete lists) to be a useful description of concepts ubiquitous to film theory and thus essential to any sort of reasoning about film and video. In addition, we point out some essential, but technically commensurable, differences between planning and story generation.

2. An Example: The "Rescue" Trailer

Our simplest video generator (which we call IDIC) uses a version of GPS [Newell and others 1963] to plan out a story; it indexes into an archive of digital video to select scenes to illustrate each part of the story generated, and then edits together the scenes into a newly created video story. The user can specify the sorts of actions that should be portrayed in the story that gets planned out by IDIC. The query to IDIC which generated the "Rescue" video represented in Figure 1 was the following:

```
(idic (gps '() '(rescue) *sttng-movie-ops*)))
```

The user calls GPS with a start state (shown as empty in the example above), a conjunct of goals, and a list of operators; then, the output of GPS is passed to IDIC which assembles the appropriate video footage together to create a new video. We have written a library of GPS operators for the domain of *Star Trek: The Next Generation* (hereafter referred to as STTNG) trailers. In other words, IDIC generates new STTNG trailers from an archive of existing trailers for STTNG episodes.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.