

# 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<sup>§</sup>

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, sit-coms, 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.

- *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,"

---

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



Negotiate



Fight



Fight (continued)



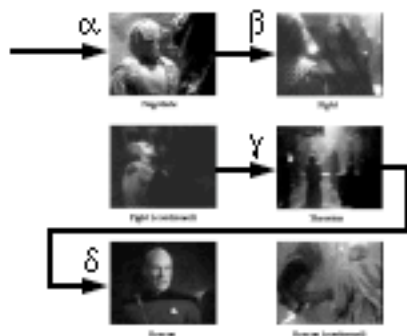
Threaten



Rescue



Rescue (continued)



$\alpha$  = establishing-negotiate

$\beta$  = break-down

$\gamma$  = threaten-renewed-violence

$\delta$  = pre-emptive-rescue

## Figure 1: The "Rescue" Trailer

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.

We are using a modified version of the GPS program that can be found in [Norvig, 1992]. Running GPS with the goal to generate a story (i.e., a story plan) which contains a rescue in it, GPS produces the following output:

```
Goal: rescue
Consider: pre-emptive-rescue
  Goal: threaten
  Consider: threaten-renewed-violence
    Goal: fight
    Consider: escalate
      Goal: threaten
      Consider: break-down
        Goal: negotiate
        Consider: appease
          Goal: threaten
```

Consider: de-escalate  
Goal: fight  
Consider: establishing-negotiate  
Action: establishing-negotiate  
Action: break-down  
Action: threaten-renewed-violence  
Action: pre-emptive-rescue

Story Plan:  
( (executing establishing-negotiate)  
(executing break-down)  
(executing threaten-renewed-violence)  
(executing pre-emptive-rescue) )

The final story plan produced uses four GPS operators (establishing-negotiate, etc.). IDIC illustrates GPS's output using four scenes selected from its archive of digital video. Figure 1 contains six stills from a video that was automatically created by IDIC in response to the query above. The stills are numbered in the temporal order of the video produced by IDIC. The lower-left corner of Figure 1 is a "road map" describing how GPS planned out the four scenes that constitute the final video. The six small panels in the lower-left are reproductions of the six stills which can be seen at the top of the figure. The connections between the scenes represented by the stills are noted in the lower left by a sequence of arrows. Each of the arrows is labeled with the GPS operator which was used to link two scenes together. Figure 1 is, thus, a representation of the video produced by IDIC and summarization of how the GPS operators "explain" the connections between the different scenes in the video.

### 3. Generating Trailers with GPS

#### 3.1 Why Star Trek The Next Generation Trailers?

We chose to analyze, represent, and generate trailers of the popular syndicated television series *Star Trek: The Next Generation* for several reasons: its characters and stories all take place within one limited, yet rich narrative universe; and there is a practice among Star Trek fans of re-editing shows as well as generating new stories within the narrative universe of Star Trek which has been studied by researchers [Jenkins 1992]. Furthermore, trailers, because of their length, are also a tractable object of study both for practical (disk space) and theoretical (narrative complexity) reasons.

#### 3.2 Representing Media: Audio and Video in STTNG Trailers

The first step in making a video generator is to analyze the structure of what is to be generated. For STTNG trailers, as with most videos, the main structural decomposition is into separate video and audio tracks. These tracks can be broken down into logical

segmentations: for the video, scenes separated by cuts; and for the audio, dialogue segments separated by pauses.

In analyzing the structure of the dialogue segments we found an unexpected result: a single STTNG trailer can be decomposed into two separate, yet coherent, trailers which have no scenes in common. For example, as shown below, by annotating a single trailer according to who is speaking in each scene, we are able to make two trailers: a trailer in which only the narrator speaks, and a trailer in which only the characters speak (the italicized text).

**Narrator:** On an all new episode of Star Trek the Next Generation...

*Duras: You are a traitor!*

**Narrator:** Worf is accused of treason and faces a Klingon death penalty.

*Worf: It is a good day to die.*

**Narrator:** His enemies are hiding the truth that could free him.

*Picard: You will not execute a member of my crew.*

**Narrator:** Now Picard must risk his life to defend Worf's innocence

**Narrator:** on the next all new episode of Star Trek the Next Generation.

These two different trailers also elucidate the relationship between the separate audio and video tracks of the movie. Listening to the audio only, the *narrator's trailer* tells a coherent story in which the characters are named (Worf, Picard, enemies) and the action and basic conflict are described. The *characters' trailer* relies far more on the video track for its coherence. Characters are identified by being seen rather than being spoken of (for they predominantly speak in the first or second person, rather than being spoken about in the third person as in the narrator's trailer), sentences often contain deictic references which can only be resolved visually (in another trailer Data says "You cannot survive in this." -- this sentence relies on the video to fill in what "this" refers to), and the main action and conflict of the story are depicted in the video rather than described in the audio. Given these results one can postulate two theoretical extremes into which a trailer can be decomposed: a trailer which is audio only and a trailer which is silent.

There are relevant historical examples for both theoretical extremes. For an audio-only narrative the examples are numerous ranging from pre-literate oral storytelling and epic poetry to radio plays. For a video only narrative there are interesting precedents from theater and film: the "dumb show" performed in *Hamlet* III.ii reminds us of the theatrical practice of pantomime stories (a "dumb show" functioned as a sort of "trailer" for a play by silently enacting its plot in short form before the play began); and the triumphs of Chaplin and

others from the early days of cinema provide us with a rich tradition of silent visual narratives (in its purest non-verbal form this corpus would encompass scenes which did not make use of "title-cards" to provide narrative information).

In the case of audio only narrative, one can imagine a trailer in which the representation of action and story is wholly dependent on a representation of the dialogue. This task would then be one of text interpretation in natural language (ignoring for the moment the use of non-speech audio in the trailer). In the case of video only narrative, the representation of action and story would rely wholly on the representation of the visual events and transitions in the video, of story elements which are intelligible without any use of sound. An ideal representation of a trailer would capture the structure and content of the audio, the video, *and* their complex interrelations. For this project, we focused on the representation of silent trailers for two reasons: 1) to avoid having to devote most of the research to natural language processing; and 2) to focus our efforts on the representation of a media type which has been largely ignored in AI research and whose impact in everyday life is enormous.

Our representation can be easily expanded to include non-speech audio because these audio events can be represented within the same scheme for the representation of visual events, either as audio reinforcements for events happening on-screen or as audio stand-ins for events happening off-screen.

### 3.3 GPS Operators and Cinematic Transitions

The most salient articulation in the video stream is the transition which links two shots. The process of linking visual shots by transitions (usually cuts) is known as "montage." For many theorists, montage is the essential feature that distinguishes cinema from other storytelling arts [Eisenstein 1949]. By applying GPS to the representation of events in video we found a surprising similarity between the structure of a GPS operator and a transition between shots. A GPS operator has a *list of preconditions*, an *add list*, and a *delete list*. For example, the definition of the GPS operator for "driving a son to school" includes the following (we have borrowed the following syntax for GPS operators from [Norvig 1992]):

```
(make-op :action
  'drive-son-to-school
  :preconds
  '(son-at-home car-works)
  :add-list
  '(son-at-school)
  :del-list
```

```
'(son-at-home))
```

The operator, drive-son-to-school, is executed by the GPS program when the preconditions, '(son-at-home car-works), are extant and the goal, '(son-at-school), is the stated goal, and when the operator defined above is a member of the set of available operators.

If one were to portray cinematically the same chain of events, the GPS operator provides a storyboard, as it were, of which information to show and which information to elide. Imagine a video sequence which establishes the preconds, '(son-at-home car-works), then cuts to the add-list, '(son-at-school). As viewers of this sequence we infer the action, 'drive-son-to-school, without having seen it. Conversely, if we only portrayed the action, 'drive-son-to-school, as, for example, a scene of a parent and son driving down the street, the preconds and add-list would be ambiguous. This could be a scene for any driving action between any two locations. So the GPS operator can encode knowledge about what we need to see and don't need to see in order to make inferences about the event structure of a video sequence. The preconds represent what needs to be shown in the first scene of a sequence, the add-list represents what needs to be shown in the next scene of a sequence, and the action of the GPS operator represents what does not need to be shown, what is supplied by the inferential activity of the viewer in the "cut" between the first scene and the next scene of the sequence. The cognitive process of the construction of video narratives through the viewer's inferential activity has been investigated by David Bordwell and his students [Bordwell 1985].

This function of the transition between shots begins to articulate some of the differences between knowledge representation of the world (common-sense) and knowledge representation of representations of the world (media).

### 3.4 GPS Operators for STTNG

In our representation of the structure of STTNG trailers we define a space of GPS operators out of only four primitive goals: *threaten*, *negotiate*, *fight*, and *rescue*. For example, the operators for "threaten renewed violence" and "pre-emptive rescue" are defined in terms of the goals fight, threaten, and rescue:

```

;fight -> threaten
  (make-op :action
    'threaten-renewed-violence
      :preconds '(fight)
      :add-list '(threaten)
      :del-list '(fight))
;threaten -> rescue
  (make-op :action
    'pre-emptive-rescue
      :preconds '(threaten)
      :add-list '(rescue)
      :del-list '(threaten))

```

As can be seen with these two operators, GPS can string together chains of operators in order to satisfy a goal. The video sequence which would result from chaining through the above operators would be a three scene sequence: open with a fight scene; cut to a threat scene; cut to a rescue scene. The GPS operators find their visual representation in the *transitions* between the scenes which are represented by the goals fight, threaten, and rescue. The “threaten renewed violence” action is inferred by the viewer in the transition from the fight scene to the threat scene, and the “pre-emptive rescue” is inferred by the viewer in the transition from the threat scene to the rescue scene.

### 3.5 Planning vs. Story Generation

Our first experiments in which we tried to use GPS to plan out a story were unsuccessful. All of the stories that were planned out by GPS using our space of operators were too short and too boring: few complications or unexpected turn of events occurred in the stories produced. It then occurred to us that there is a very important difference between plans and stories: in plans one normally values the short and simple, while in stories it is the unexpected and complicated events which one looks for. We made one important change to GPS in order to coax it to produce stories instead of plans: we programmed GPS to select the operator having the *most* unsolved preconditions rather than, as is usual for planning, to select the operator with the *least* number of unsolved preconditions. This change, which involved changing one character in the program (specifying a sort function to use > as an ordering function rather than <), allowed GPS to produce better stories.

## 4. Current and Future Work

Our system, IDIC, can generate a coherent trailer using GPS-like mechanisms and representations. However, we feel that the success of our system is largely a result of the fact that we are only linking one video segment to each GPS goal. This approach manages the complexity of the task of video representation and story generation by not explicitly representing all the knowledge we ourselves used in selecting the

appropriate video segment to match each GPS goal (threaten, fight, negotiate, and rescue).

In our current work in story generation we are using a much richer set of representation languages to describe the content of the video. In some television, for example, news broadcasts, the dialogue provides the backbone of the story. Consequently, our current work has expanded to include the use of various natural language processing (NLP) techniques to analyze and represent dialogue (e.g., [Sack 1993a], [Sack 1993b], [Haase, 1991]). In the near-future we intend to apply these NLP techniques to the automatic analysis of the closed-caption text that is often incorporated into television broadcasts. Segments from the television broadcasts analyzed with the NLP techniques could then be used as an archive of video segments for a story generator.

Another line of our research is a detailed examination of sound and picture in film and video. This research builds on top of and will be expanded within a prototype system being developed at the MIT Media Laboratory, *Media Streams* [Davis, 1993] which uses an iconic visual language to create temporally indexed, multi-layered content annotations to describe many of the aspects of video that a system would need to represent in order to retrieve and repurpose segments of a video stream: spatial location, temporal location, weather, characters, characters' actions, characters' relative positions, characters' screen positions, objects, objects' actions, objects' relative positions, objects' screen positions, camera motion, camera framing, shot breaks, and transitions between shots.

*Media Streams* makes use of *FRAMER*, which is a recursive persistent framework for knowledge representation and media annotation developed by Prof. Ken Haase at the MIT Media Laboratory (see [Haase 1993] and [Haase 1994]). The categorization of shot transitions used in *Media Streams* is based on Noël Burch's taxonomy [Burch 1969] which was used by Gilles Bloch in his work on video representation [Bloch 1987].

Creating a video generation architecture which can make use of detailed content annotations will necessitate moving beyond a simple planner in two ways:

(1) *Matching Goals to Video Segments*: In our IDIC system we have assumed that matching between GPS goals and video segments is a simple process. However, when video segments are described using multiple annotations matching is no longer simple. Matching becomes one of the main pieces of work that the system must accomplish. We are currently investigating the use of more complex, analogical matching techniques.

(2) *Goal Interactions*: With the introduction of multiple annotations, and a more complex matching routine which allows multiple video segments to be matched to a single goal, one must consider the problem of how the "best" match can be found. The "best" match for a given goal can only be computed by taking into consideration its interaction with the "best" matches for other goals. Processes for computing the "best" matches for sets of goals will necessitate extending the architecture to include mechanisms responsible for the management of narrative continuity and complexity.

## 5. Conclusions

We have described 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 makes use of an old model of planning [GPS, Newell and others 1963]. By incorporating a re-implementation of GPS into IDIC we have been able to explore the various ways in which the mechanisms embodied in GPS are and are not appropriate to the task of video generation. IDIC is one of a set of video generation systems we are building to "review" the literature of artificial intelligence in search of theories and techniques for the description of the form and content of television, one of the most ubiquitous sources of information (i.e., sources of "common sense") in the contemporary United States.

## Acknowledgments

The research discussed above was conducted at the MIT Media Laboratory where we are members of the Machine Understanding Group in the Learning and Common Sense Section. The support of the Laboratory and its sponsors is gratefully acknowledged.

## References

- Bloch, Gilles R. "From Concepts to Film Sequences." Yale University Department of Computer Science, 1987.
- Bordwell, David. Narration in the Fiction Film. Madison: University of Wisconsin Press, 1985.
- Burch, Noël. Theory of Film Practice. Translated by Helen R. Lane. Princeton: Princeton University Press, 1969.
- Chapman, David. "Planning for Conjunctive Goals." Artificial Intelligence (32 1987): 333-377.
- Cross, Donna Woolfolk. Mediaspeak: How Television Makes Up Your Mind. New York: Penguin Books, 1983.
- Davis, Marc. "Media Streams: An Iconic Visual Language for Video Annotation." In: Proceedings of 1993 IEEE Symposium on Visual Languages in Bergen, Norway, IEEE Computer Society Press, 196-202, 1993. (extended version in: Teletronikk 4.93 (1993): 59-71.)
- Eisenstein, S. Film Form: Essays in Film Theory. Edited and Translated by Jay Leyda. New York: Harcourt Brace Jovanovich, Publishers, 1949.
- Haase, K. "Making clouds from cement: Building abstractions out of concrete examples" In Proceedings of the US-Japan Workshop on Integrated Comprehension and Generation in Perceptually Grounded Environments, 1991.
- Haase, Ken. "FRAMER: A Persistent, Portable, Representation Library." Internal Document. Cambridge, Massachusetts: MIT Media Laboratory, 1993.
- Haase, Ken "FRAMER Manual." Internal Document. Cambridge, Massachusetts: MIT Media Laboratory, 1994.
- Hobbs, Jerry R. and Moore, Robert C., eds., Formal Theories of the Commonsense World. Norwood, NJ: Ablex Publishing Corporation, 1985.
- Jenkins, Henry. Textual Poachers: Television Fans & Participatory Culture. Studies in Culture and Communication, ed. John Fiske. New York: Routledge, 1992.
- Lenat, Douglas B. and Guha, R. V., Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project. Reading, MA: Addison-Wesley Publishing Company, Inc., 1990.
- McCarthy, John. "Programs with Common Sense." In Proceedings of the Symposium on the Mechanization of Thought Processes, National Physical Laboratory, Teddington, England, 1958.
- Newell, Alan, J. C. Shaw, and Herbert A. Simon. "GPS, A Program That Simulates Human Thought." In Computers and Thought, ed. Edward A. Feigenbaum and Julian Feldman. 279-293. New York: McGraw-Hill, 1963.
- Norvig, Peter. Paradigms of Artificial Intelligence Programming: Case Studies in Common Lisp. San Mateo, California: Morgan Kaufmann Publishers, 1992.
- Sack, W. "Coding News and Popular Culture," paper presented at the International Joint Conference on Artificial Intelligence, Workshop on Models of Teaching and Models of Learning, R. Schank (organizer), Chambéry, France, 1993a.
- Sack, W. "Recognizing Rhetoric in News Stories" Master's Thesis Proposal. Cambridge, MA: MIT Media Lab, 1993b.