2. Background: Representation and Interface Design

This chapter introduces the initial problem of manipulating video in a computer environment, and presents background information about artificial intelligence, learning theory, and visual interface design.

2.1 Editing Images

Part of the pleasure of home video -- of, in fact, any kind of image making -- is in shooting, the act of creating the image. Another part is in viewing the image and sharing it with others.

For a first viewing, one watches everything that was shot. After vacation, the family may gather together to view the video tapes; at the end of the day, a professional film crew watches the "dailies" [PA 84]; the amateur photographer may flip through a stack of snapshots in the parking lot outside of the photo store.

Subsequent viewings benefit from editing. Not all the film shot is included in the final movie, and not all the snapshots taken are placed in the scrapbook. Home video is not often edited, but would benefit from this process. Reducing the amount of material makes repeat viewings more enjoyable. One can shorten the time required to view the images, improve the average quality by selecting pleasing ones, and create new meanings by juxtaposing them. The editing process is also an end in itself; for example, the act of making a scrapbook is another way to enjoy the snapshots.

2.2 The Search Problem

In order to view a remembered chunk of video, it is necessary to *find* it. Professional editors spend a significant portion of their time describing images so that later they can quickly find particular film clips. The ability to find images easily encourages their creative use.

To find a chunk of video, it is necessary to have a representation of the material available. There are many different kinds of representations. The camera-person's memory of what was shot is a representation. The label placed on the tape (for example, "Italy 1989") is a representation designed to jog the memory of a person -- the camera-person or others with an interest in the event. The camera-person's memory includes context information which may not appear in the video itself -- where and when it was shot, what happened before and after, and the like. Human memory is a powerful

representation system, and many other representations are used as aids to human memory.

A description of the content of a video tape is a called a "log." The primary purpose of a log, like any other representation, is to help the user to retrieve the material later. A paper description helps the user to perform the action of retrieving the desired image by telling him/her where to look. A computer description allows the system to find the image, saving the user work.

2.3 Computer Representations

The simplest form of computer representation is to divide video into segments[1] and give each segment a title [Sasnett 86]. Title selection is idiosyncratic. One person's titles may not be meaningful to another, or even to the original person after time has passed. Finding a desired segment or type of segment from a list of titles can be difficult.

Keywords make it easier for a person to find segments with the aid of a computer. This approach is used in *New Orleans in Transition, 1983-1986* [Davenport 87] and in *The Elastic Charles* [BD 89], multimedia documents developed at the MIT Media Lab. In The Elastic Charles, a video segment of a man fishing in the Charles River locks is marked by the keywords "locks," "recreation," and "fish." One could search for other instances of each of these keywords.

This is much more useful than searching through the text of segment titles, but it has limitations. For example, searching for all instances of the keyword "toy" will not return segments with a "stuffed animal" or "bubble stuff." The system does not know that a stuffed animal is a kind of toy. It also does not know that "stuffed animal," "bubble stuff," and "fork" are all objects. The system may use the keyword "parent," but there is no representation for the fact that "parent" is a relationship between two people, or that it has an inverse relationship, "child."

---

[1] A note on terminology: the word "shot" is part of traditional filmmaking terminology, and is used in two ways. Literally speaking, a shot is the contiguous series of frames from the moment when the camera was turned on to when it was turned off. However, the term "shot" is also used to refer to a set of frames that capture an action.

The word "sequence" also has two meanings. Literally speaking, a sequence is any set of shots combined together. A sequence is also the set of shots which capture an event. Thus, one may refer to the "chase sequence " of a movie.

The term "segment" has come into use in the field of computer manipulation of video. A segment is any set of frames between an in-point and an out-point on a video source. Therefore, a segment could contain a shot or a sequence in all of the senses described above.

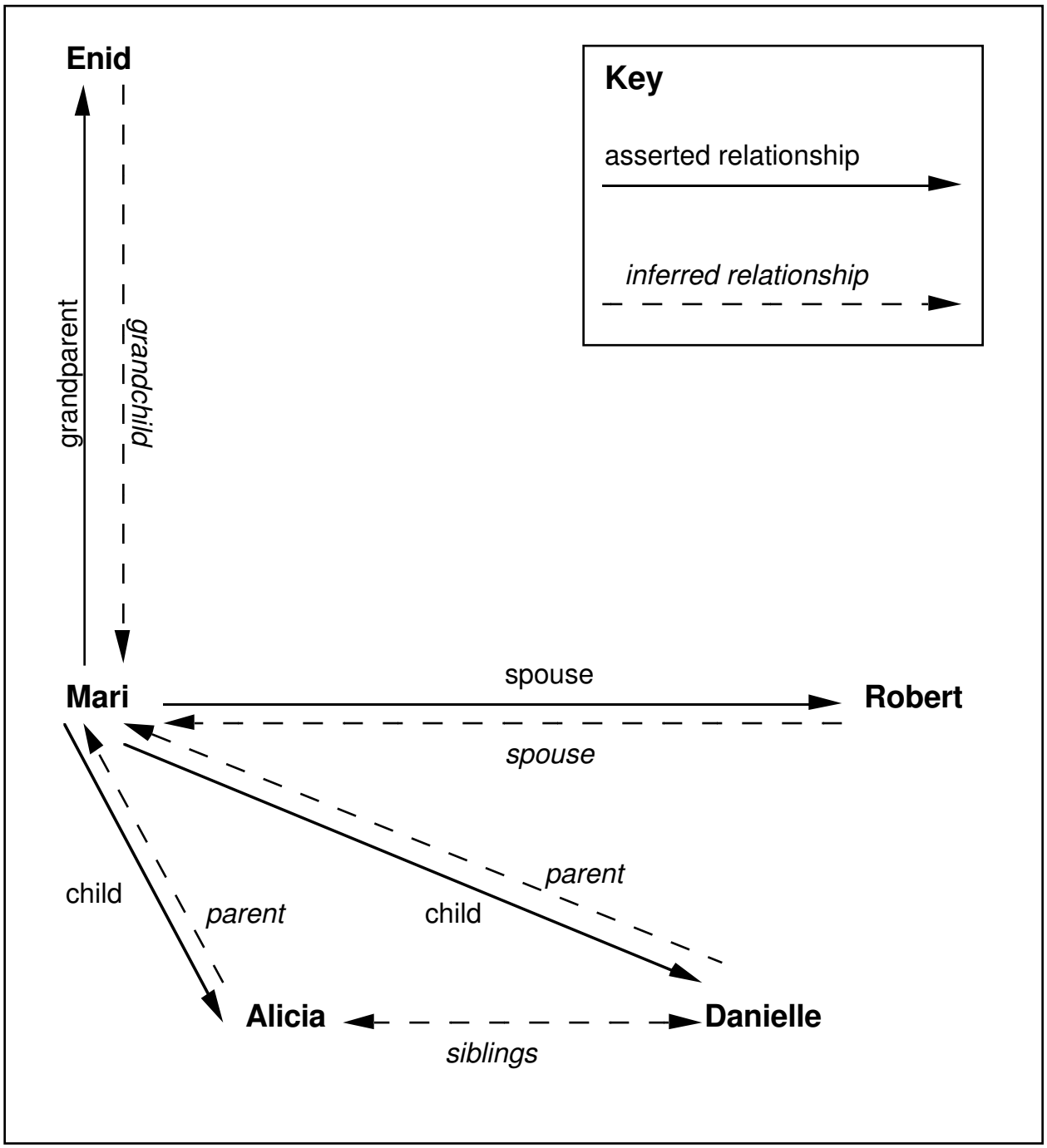

Figure 2.1: A semantic family tree. Mari's relationship to each other person has been asserted by the user (solid lines), and the computer infers the inverse relationships and the relationship between the children (dotted lines). All that is needed to complete the diagram is to assert Robert's relationship to the children. A total of six asserted relationships leads to fourteen inferred relationships.

Semantic knowledge representation[2] can help solve these problems. The Electronic Scrapbook uses **Arlotje**, a knowledge representation system developed by Kenneth Haase [Haase 90]. Searching for all "toys" will return balls, bats, stuffed animals, bubble stuff, and the like, because these objects are defined as subsets of the collection[3] of all toys. Toys in turn are subsets of the collection of all objects. A search for all objects will return forks, knives, and sofas as well as toys.

Properties of units and relationships between units can be tracked. Instead of "Alicia" being simply a keyword, Alicia is an instance of humans with slots that record her birthday and her relationship to other humans.

Inferencing mechanisms help to maintain relationships between units. For example, if we know that Mari is Alicia's mother, then we know that the inverse relationship holds -- Alicia is Mari's child. If it was already known that Danielle is Mari's child, then we know that Alicia and Danielle are either whole or half sisters (see Figure 2.1). Making these inferences saves the user time. Furthermore, family relationships are useful abstractions around which to structure stories.

2.4 How Much Representation is Necessary?

How much representation is necessary? Consider the issue of objects. Is it necessary to represent every object visible in the frame? Is this information useful? A complete description of the objects surrounding me as I write this would fill a small volume: the clutter of books, papers, computer disks, furniture, and photographs. Does it matter that my desk was purchased unfinished four years ago at Market Square Hardware, was painstakingly sanded and polyurethaned by me in the backyard with two coats to the front and four to the top, and that there's a small dent in the back-left corner where it was dropped when I moved? How much detail is useful? The answers to these questions depend on what the representation will be used for. There is no such thing as a completely general representation.

In this particular example, the objects which seem most significant to me are my computer, my desk (extraneous detail omitted), and my cup of tea. These objects play a functional role in the scene. If one were making a home video about me, this would seem to be an appropriate level of detail. This is a subjective judgement.

A representation must be evaluated in terms of its function: does it enable the system to perform the task at hand? What kinds of new tasks might the

---

[2] A good basic reference on semantic networks is [Winston 84].

[3] For an introduction to the terminology used in Arlotje, see section 3.2 or [Haase 90].

information be used for in the future? Is the representation robust enough to be used beyond its immediate purpose?

2.5 Interface Design and Coding

The primary, original motivation for this thesis was to investigate how to represent knowledge about video. Strategies for knowledge representation, their benefits, and their shortcomings will be discussed throughout this document. As programming work began, however, I discovered that for every 10 minutes of knowledge representation work, there was 10 hours of interface coding.

No information can be used unless it can somehow be entered, stored, retrieved, displayed, and edited. How these tasks can be performed simply and efficiently are not peripheral issues. There is a direct cost/benefit trade-off between the time it takes to do knowledge representation and the tasks the information allows the user to perform. One must also account for the time necessary to learn to use the system; it should not require a computer science degree to edit a video. How can abstract concepts like knowledge representation, electronic signals, and narratives be made understandable to the naive user?

Consequently, this thesis became as much about interface design as about knowledge representation. In this area, the work is inspired by research being done at the Media Laboratory in the Learning and Epistemology Group and the Visible Language Workshop.

2.6 Constructionism

In The Electronic Scrapbook, video is divided into segments which are treated as objects. Seymour Papert's Constructionist theory of learning emphasizes the importance of "objects to think with" [Papert 80]. Treating segments of video as objects is a powerful metaphor. One can make a video segment, put it somewhere, describe it, copy it, or throw it away. These metaphors are clearer to the average user than the concept that one can create a pointer to a section of an electronic signal, put the visual representation of that pointer in a window, create a data structure to be used as an index to retrieve the pointer, make multiple pointers to the same section of video signal, and discard pointers which are no longer needed.

This work is also influenced by the Learning and Epistemology Group's vision of computers as tools to help people learn. It is hoped that through using the system, people will gradually learn about the medium and its potential. The system is designed to promote this process. The typical home

video artist does not need a handbook of conventions, but , rather, an environment which encourages him/her to experiment with the medium.

In my experience as a pinball player, I have observed that you can become good at any game that you enjoy losing. In Papert's terms, it is important not to ignore affective properties; learning can and should be fun. I hope that people will enjoy using The Electronic Scrapbook and that it will help them to learn by doing.

In the design of The Electronic Scrapbook, the human and the machine are viewed as partners; each enhances the abilities of the other. Work in artificial intelligence too often tries to create "Einstein in a box" [Hewitt 91] -- systems which do work for people, rather than working with people.

2.7 Visual Interface Design

I began this project with a commitment to making the interface as visual as possible. As the reader will discover, I learned a number of lessons about things that text can communicate more clearly than graphics. A healthy combination of modalities is preferable; each medium should be used for what it does best.

The idea that a medium should be used *only* for what it does best spawned the Modernist movement in art. Post-Modern theory argues that one must try to understand the potential of each medium thoroughly, but feel free to combine them as one wishes.[4] A successful work will communicate with different people on different levels. The difficulty in applying these art-historical ideas to human-computer interface design is that one must work with a more rigorous definition of "communication." In art, communication is subjective; in interface design, it is closer to objective: can the user figure out how to perform the task?

Visual communication is in many ways more challenging than written. Most people spend years learning how to write, but few ever think seriously about visual communication. Certainly, textual communication is preferable to poorly-designed graphical communication. While I do believe that I have developed some insights into the uses of these different modalities, I wish to

[4]The critic Clement Greenberg wrote that each medium should limit itself to what is unique to that medium [Greenberg 61]. Thus, painting should eliminate representation of the real world, because that is the realm of literature. Furthermore, painting should also eliminate representation of the third dimension, because that is the realm of sculpture. Greenberg's writings were influential and helped to spawn movements in abstract art, especially Minimalism. Contemporary critics such as Timothy J. Clarke argue that while a great deal was learned about the properties of various media through these rigid restrictions, the time has now come to discard them and apply the knowledge gained [Clarke 87].

state up front that I am not a trained graphic designer. Research in visual interface design would benefit from greater collaboration between interface designers and graphic designers.