Let's See That Again: A Multiuse Video Database Project

by
Lee H. Morgenroth and Glorianna Davenport
Interactive Cinema Group
MIT Media laboratory

Abstract

In the past, video production has had three distinct phases: content collection, logging, and video editing. Production was a fairly linear process, with little overlap or communication among the three phases. Typically, logging supported editing to produce single-use structures. With the advent of digital video technology, we can imagine the production of multiuse video databases. These databases pose the problem of how to describe content for multiple uses.

In addition to the problems involved in building video databases, users face major problems in navigating the database. Users may have trouble locating their desired content because they don't know what is available, they don't know how to effectively use the modes of access, and they, perhaps most importantly, don't know what they want.

The video database project discussed in this paper is rethinking the role and use of description in video databases. A toolset of three applications has been created to give the user more powerful control over video. These applications also aim to make the process of attaching and using descriptions for video more efficient. These tools minimize the task of conventional annotation, and redirect that energy to the process of making stories. By using the toolset for creating stories, the user encodes story-based annotations and expert knowledge about editing into the database. Thereby, the database grows and becomes structured by the process of building stories. This structuring process optimizes the database for retrieval of video in a story form.

Introduction

Movies are expensive. We have all heard the figures out of Hollywood for the production cost of the latest blockbuster. But even low-budget feature films run production costs into the millions. In a lo- budget film, the production costs of a single frame of film might be seven dollars. One second of screen time would cost over $150. In a big-budget film the average cost of a single frame can be over $500. These figures bring the cost of the average second of Arnold Schwarzenegger to $16,000. Of course the economics of a film cannot be divided equally between all the frames in the film. Some sequences cost substantially more than others. So, the big scenes in the big films can cost hundreds of thousands of dollars to produce. But the most amazing thing about the cost of movies is that all this ultra-expensive footage is used only once. The million-dollar scene has its place in the movie, and that is the only place you will see it.

Motion pictures have traditionally been a single-use medium. Generally, images are not used in more than one movie, or perhaps more surprisingly not more than one commercial. Even if the director can specify the need for a shot of "woods turning color in the fall," the probability of finding and buying the "perfect" image at a reasonable price and cutting it into a highly structured commercial is remote. There are probably more constraints on the shot than have been specified. If a budget exists, it is easier and cheaper to design a new shot.

Nonetheless, over the years the idea of reusable video has generated some interest. In the past, stock footage houses collected both generic scenes and special event footage. Today, Kodak is building a system which allows remote access to a large picture database. Many other enterprises are looking at the potential of using advanced network technology to provide access to stills, motion pictures, and sounds. At times the fanfare surrounding the development of these advanced technologies obscure the significant issue of the production of appropriate and desired content. A few years back, a group at the MIT Media Laboratory experimented with repurposing footage from the soap opera *Dallas* for interactive replay. The experiment failed because this apparently multithreaded soap opera was too tightly structured to be repurposed. The obvious lesson of this experiment was that if we are to build multiuse databases, we must think about a story purpose and then create content for that purpose.

A current project in future image services at the MIT Media Laboratory proposes a model for building multiuse video databases which begins with content (Pentland, 1993). The collection of content is supported by tools which combine story generation and video annotation. We believe our premise of story-based interaction will provide an important foundation for future navigation in large video databases.

Content and Tools

A New Application

Developing a multiple-use database of video involves adding constraints to the already structured task of video production. The choice of content is one area that becomes more constrained. The *Dallas* project showed that fictional dramas are a difficult domain for video reuse. A better choice of content may be travel.

Travel planning can be an involved experience. People routinely read brochures and articles, talk with friends and travel agents, and even watch movies to help them decide where to go on vacation. Because people are willing to and even enjoy putting effort into travel planning, a video database designed to aid potential travelers is an appropriate service. The high entertainment value of video matches the excitement and feeling of adventure that surrounds travel. A database that contains video profiling an array of travel destinations would be a useful tool that travel agents could use to provide their customers with visualizations. Some travel agents already maintain libraries of videotapes on various locales. A video travel database affords at least one feature that a library of videos does not -- reconfigurable shot selection for the purpose of personalization.

The vision for the Travel Database of the Future is one in which a travel agent can generate a story on the specific destination that fits the client's tastes. So, the traveler considering a trip to England who is interested in theater would be able to view a video about the West End of London and Shakespeare's home in Stratford-upon-Avon. Whereas a traveler with a pension for sailing might make travel decisions based on video of some of the many coastal towns on the eastern shores of England.

There is a large amount of work involved in developing the tools and content to provide such a service. Development of most of the tools is a one-time cost, whereas content collection is an ongoing task. A better understanding of the relationship between the tools and the content in the video database system can make the production process more efficient.

A New Toolset

A toolset consisting of a database browsing tool, a story generation tool, and a visual editing tool has been implemented in the Interactive Cinema Group at the MIT Media Laboratory. The toolset was implemented in a UNIX environment using Motif for interface design. The three tools share a common data representation. The representation is implemented in the Framer knowledge representation language developed by Prof. Ken Haase in the Learning and Common Sense section of the MIT Media Laboratory (Haase, 1993). The common data representation allows the three applications to share data generated by each. These three tools in communication create a database system that can be evolved

through use. To test these tools using appropriate content, video was shot to create a prototype video travel database.

Example Content

The video used to create the prototype database was shot in Woodbridge, England. The footage was collected in July 1993 over a shooting period of two weeks. A total of four hours of material was collected, using a commercial Hi-8 camcorder. The raw footage was culled down to one hour of unedited source material to be used in the video database. The most important aspect of the content collection process was that the footage was shot specifically for a multiple-use database. This intention manifested itself in all aspects of the video. Overlapping stories about several aspects of the town of Woodbridge made up the bulk of the content.

Three main themes emerged from the footage of Woodbridge. These themes were: the history of Woodbridge, the river Deben and its relationship with Woodbridge, and the pubs of Woodbridge. Following the culling process, the video has to be transformed through annotation into a database.

Annotation

If I have ten minutes of interview footage of the former mayor of Woodbridge, how do I make this usable in database form? If the machine is expected to sequence the video into a story, some information about the video must be added to the database. The database must know about the content of the video and how to segment it. In 1992, Thomas Smith developed an annotation system for video at the MIT Media Laboratory called Stratification (Smith, 1992). This system of annotation serves as a base for this multiuse video database project. The Stratification system treats the video as an uninterrupted stream of frames. The annotation tool allows descriptions to be attached to any group of contiguous frames. The Stratification system also allows for layering of descriptions on the video stream. Using layered description, any group of frames may have a number of descriptions associated with it. This allows the video to be described at different granularities and in different contexts (Goldman-Segall, 1993).

The stream-based style of logging used in Stratification affords some advantages over clip-based logging. In clip-based logging, the video is segmented into discrete clips. Descriptions are then attached to the entire clip. If two descriptions are applicable, both descriptions must be attached or two separate clips must be made. In stream-based logging, the video is treated as an uninterrupted stream of frames. Annotations are then placed along this stream at various in and out points. The in and out points of one annotation are free to overlap with the in and out points of any other annotation. Annotation in this style yields patterns in the description that highlight important events in the video. These events maybe cuts, movement of characters, or changes in subject. In stream-based annotation, each description represents one possible segmentation of the video. In addition, the intersection of overlapping annotations can yield further meaningful segmentations.

Building a Story

Perhaps the best way to understand the processes and tools involved in this project is through an example. The following is an example of how one editor would use the toolset to create an introductory tour of Woodbridge.

The Stratagraph: Browsing the Database

The first task is to become familiar with the database of available footage on Woodbridge. The Stratagraph, a database browsing tool, was designed for this purpose.

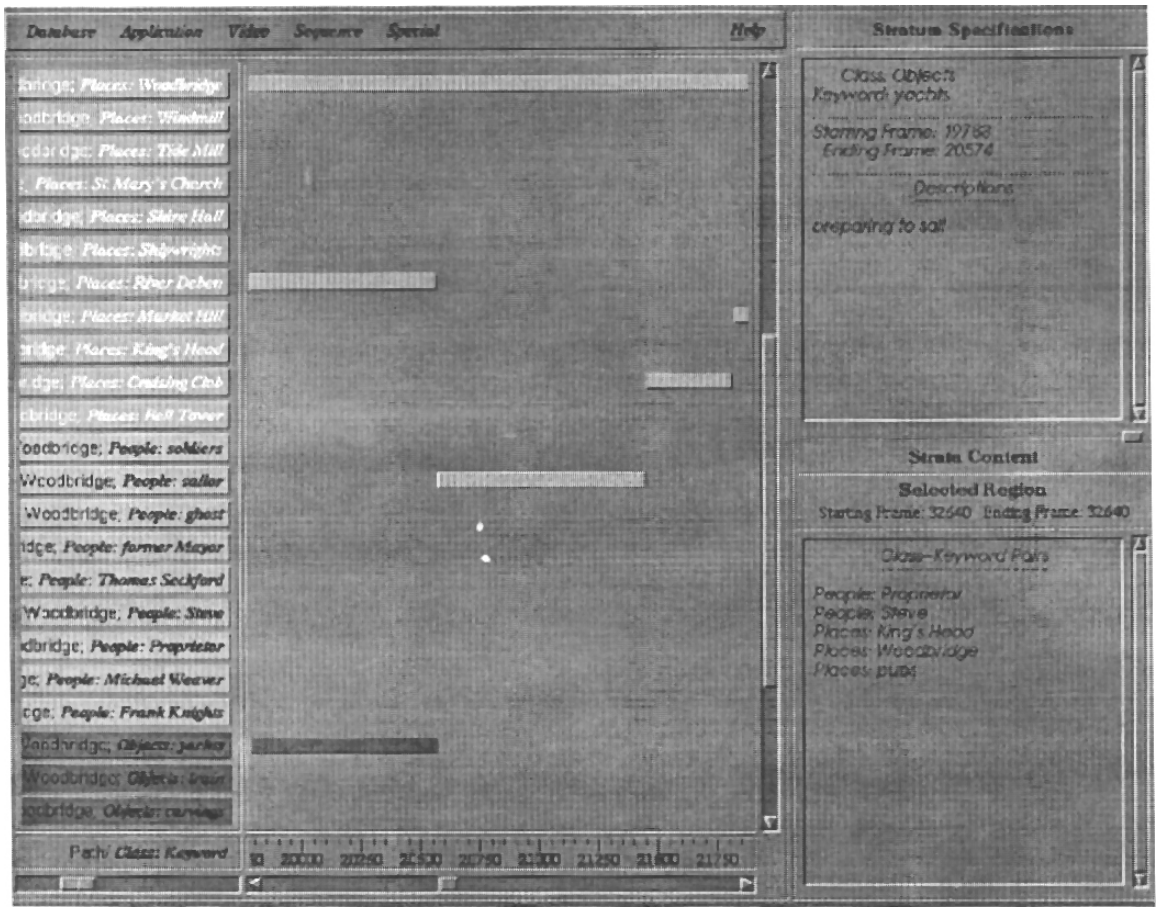

Figure 1: The Stratagraph, the database browsing tool.

The Stratagraph is a tool that allows a user to familiarize themselves with a database of annotated video. As shown in figure 1, the Stratagraph is a graphical representation of annotations in the video database. Along the y-axis of the display is a list of all the unique descriptions in the database. Along the x-axis of the display is a timeline depicting frame numbers in the video. In the main region of the display are several bars that represent actual descriptions attached

to video. These descriptions are called Strata Lines or Stratum. Each Stratum has an in-frame and an out-frame that relate to the video timeline. The in and out points determine the duration of the description, which is reflected in the length of the Stratum in the display.

The Stratagraph can be used to browse the database in several ways. One can browse by description by selecting any of the descriptions from the list in the display. When a description is selected, the display shifts to the region of the graph where the first Stratum with that description lies. The video that this Stratum overlaps can be viewed, and the other descriptions that either overlap or lie close to the chosen Stratum are visible in the display. In this way one can get not only an understanding of the selected description, but of the other descriptions for the associated video, and the context they provide. The descriptions and associated video of any other Stratum in the display can be viewed by a single mouse click.

The video stream itself can also be used as a browsing framework. Any region of the video can be selected and its description displayed. In the same way as the description-based search, the Stratagraph shows the descriptions that overlap and surround a segment of video, giving a better understanding for the context of the segment.

Homer: The Story Building Tool

Once the extent of the video available and how to access it through descriptions is understood, the story-building process can begin. Homer is the tool that allows the user to build stories from the database (Morgenroth, 1992). Homer was designed as a graphical workspace in which editors could build structures that are accurate models of the stories they wish to tell in video.

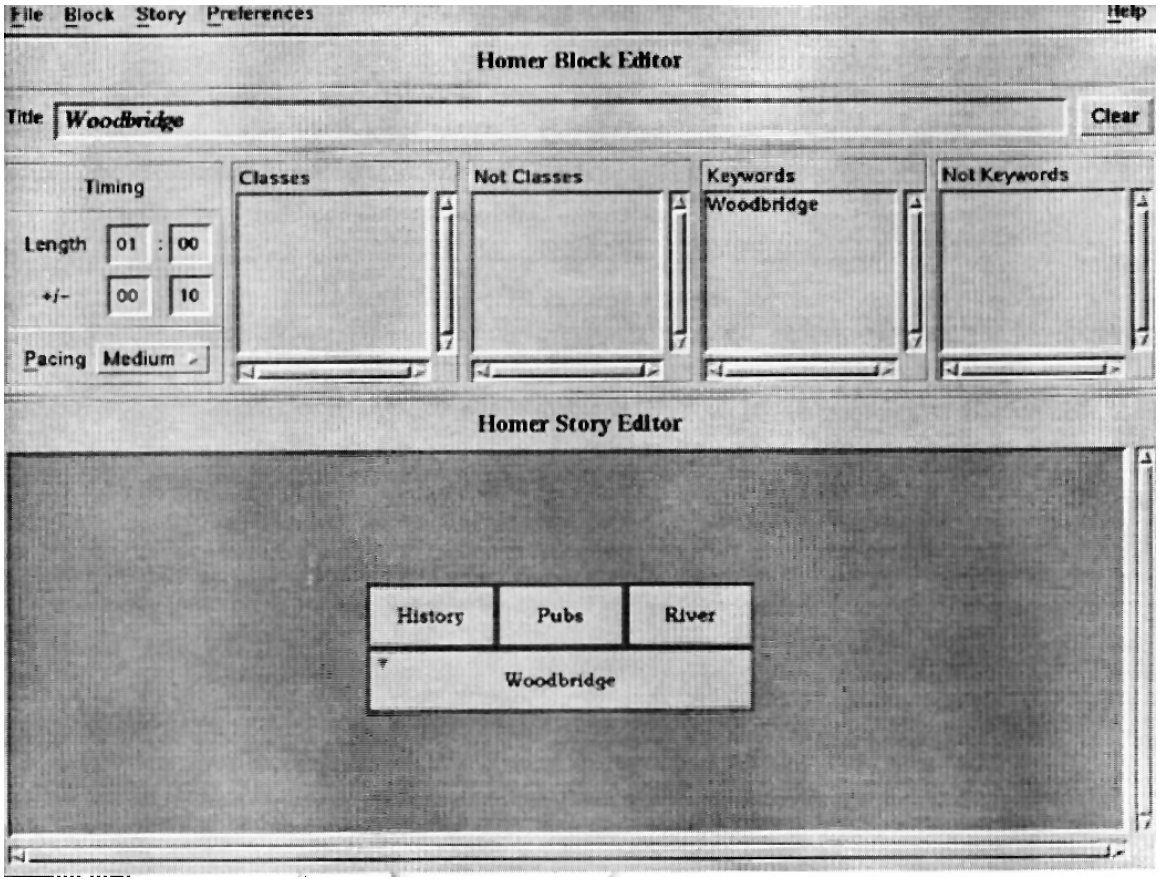

Figure 2: Homer the story model building tool.

Figure 2 shows the Homer interface and a story model. Stories are built in Homer using abstract story chunks, called Blocks. Each Block has a size, which is proportional to the length of story time that the Block covers. Block sizes can range from one second to several hours. Each Block also has a number of descriptions that determine the story content. The maker can design a story by creating a progression of Blocks. Blocks can also be layered to create sequence structures.

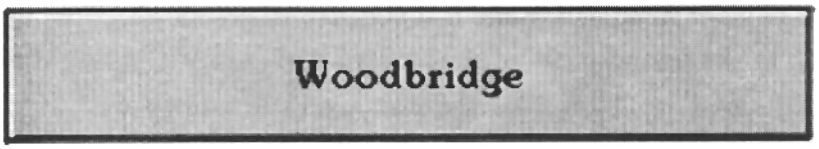

Figure 3: A single-Block story model.

The first step in building a story about Woodbridge is to create a single Block model (figure 3). At this early phase of story development, the most important aspect of the story model is the description. This single Block model has only one description: the keyword "Woodbridge." When this model is applied to the database, the resulting edit will be a mixed bag of footage about Woodbridge. The result of the first several edits generated from the single Block model contained mostly clips about the river and a few shots of the pub. Each time Homer renders an edit from a story model, it applies a weighted randomizing function to the output. This will vary the footage that is chosen for an edit, while still maintaining the constraints in the story model. So the early phases of story making are yet another style of browsing.

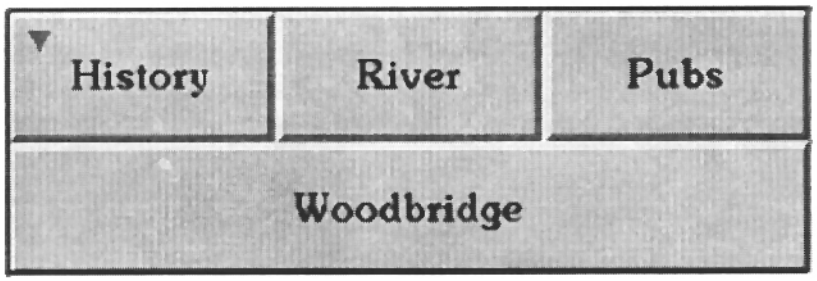

Figure 4: A story model with three sequences.

Once the type of footage the initial model produces is known, structure can be added to the story. To create an overview of the town the story can be separated into three sections, one corresponding to each major theme in the footage. As shown in figure 4, the story begins with history of the town, followed by a section on the river, and ending at the pub. As more structure is added to the story, and the resulting footage is viewed, the story model can be tuned through changes in description and timing. By using both Homer and Stratagraph together, the type of footage for each sequence can be specified. For example, if Homer came up with a nice segment of an interview with a historian, the descriptions for that segment can be found in the Stratagraph. Then, those descriptions can be added to the Block in Homer where that segment should appear. In the first phase of story creation, Homer is supplied with general descriptions, and it returns a range of footage. In the second phase of story creation, the story is broken down into sequences and the model is focused around the footage that best fits the story.

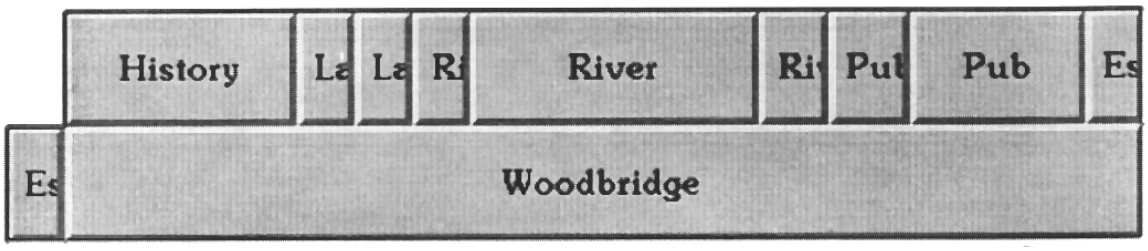

Figure 5: A story model with transitions.

Once the general sequences of the story are decided, transitions are added. Figure 5 shows a later story model that contains a number of smaller Blocks that specify transitions. The first small Block in the model represents an establishing shot for the entire story. Some of the small Blocks between the sequences are establishing shots for the sequences they precede. Other small Blocks specify shots of landmarks and other exteriors that can serve to ease the change of location in the story. This third phase of story creation brings the editing concept of a transition into the model. The final phase of story creation using Homer focuses on additional aspects of editing.

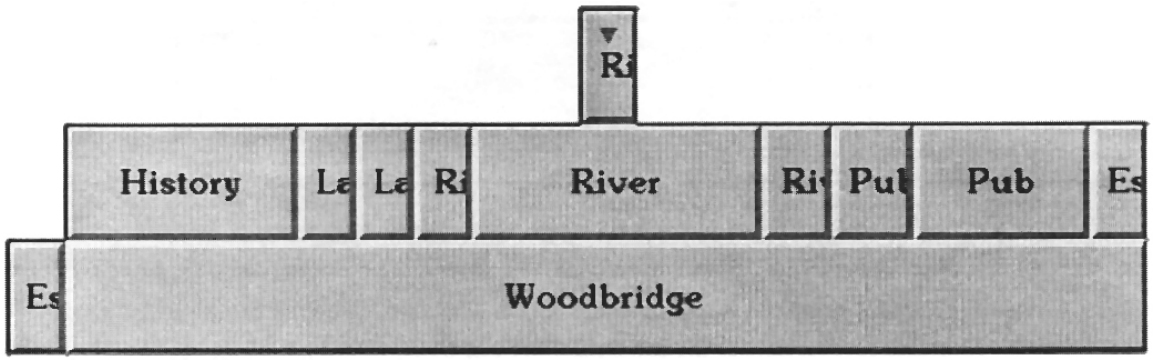

Figure 6: A complete story model with a cutaway Block.

In the stories that Homer renders from the various models, there are some shots and cuts that work, and some that do not. As the model becomes more specified, more bad shots and cuts are weeded out. Eventually the state is reached where new smaller and smaller Blocks map directly to single shots in the edit. When the description of these small Blocks is changed, this change is reflected in new shots that will fill the Blocks. The model in figure 6 has one additional Block on the top level of the model. This Block is an example of a single shot Block. It represents a cutaway in an interview sequence. The "River" Block below the cutaway Block generates footage from an interview with the former pilot of Woodbridge port. In an earlier edit generated from the model, the River Block was filled with two shots from this interview. Although the content was good, the framing on both shots were very similar, resulting in a jump cut when the two were played sequentially. To avoid this jarring cut, the cutaway Block was added in the middle of this interview. The cutaway Block constrains the section of memory that it covers, by excluding all shots of the interview with the pilot. Since that

section of the story is also described as the river, the cutaway Block is filled with an exterior shot of the river. This shot works nicely to eliminate the jump cut and still maintain continuity in the interview.

The Sequencer: Polishing the Product

At this point, although the Block model gives an accurate description of the story to be told, what is generated by Homer would be considered a rough cut. The rendered edit has most of the footage necessary to tell the story, but there are still some poor cuts and a few changes to be made. The Sequencer, the third tool in the toolset, can be used to fine-tune the rough edit. Using the Sequencer, shots can be reordered, or replaced, and cuts can be trimmed to provide better transitions. The Sequencer can be used to create a polished edit of the Woodbridge story that was built using Homer.

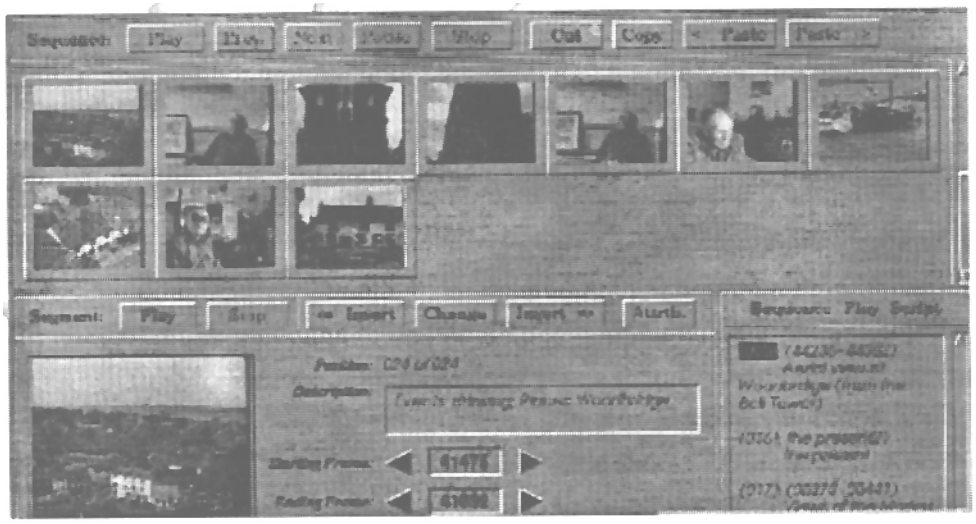

Figure 7: The Sequencer, a limited editing tool.

Closing the Loop

By using the Stratagraph, Homer, and the Sequencer, we have successfully created an edit from a database of footage. We used knowledge of story and the principles of editing to achieve this success. Since we used the toolset for the entire process, it is possible to record the decisions we made and to incorporate that information into the system. Both Homer and the Sequencer contain records of these decisions.

During the story making process, the models built with Homer are seen as a description by which Homer can select the appropriate video to create a story. But once the edited story has been created, the Block model can also be seen as a log of the content contained in the edited story. In other words, the Blocks in the Homer model supply a detailed description of the video in the story that is generated. By thinking about Homer models in this slightly different way, we can use Homer to convert the editing decisions made in the story generation process into data that can be merged with the video database.

If the finished edit from the Sequencer is re-mapped onto the Homer model, new database descriptions can be created. These descriptions combine the story information contained in the Homer model with the editing decisions made in the Sequencer. So, all the descriptions that are part of the Homer model can be automatically attached to the footage selected for the edit. This process also records the segmentations chosen in the Sequencer into the database. Through the addition of this information, the database evolves and is structured around story. Once this feedback mechanism is engaged, the system can be used in various other ways.

An Alternative to Conventional Annotation

In this prototype database, the video was annotated by hand. The inefficiency of hand logging is a problem in developing video databases. Although the Stratification system allows complex annotations to be attached simply, logging is still a burden. Fully logging a piece of video can take several times the length of that video to complete. There is some related research into image analysis in the Perceptual Computing section at the MIT Media Laboratory (A. Pentland, 1994). We hope to eventually move the majority of the logging task to an automated system based on this research. At present, machine vision technology is not at a point where it can automatically generate annotations that are detailed enough for a video database. But in combination with the tools discussed in the previous section. The bulk of the annotation process can be shifted into the editing phase.

By closing the loop on the story generation process, we can begin to use Homer as an annotation tool as well as a story making tool. As an annotation tool, Homer can be used to attach new descriptions to segments of video from the database. When creating a story model, the maker uses descriptions from the

database to determine which segments of video will be selected for each region of the story. When using Homer for annotation, the maker should also include in the Block model descriptions of how the video will be used in the story, regardless of whether or not these descriptions exist in the database. Then, once the maker has created a successful story from the model, the new descriptions can be attached t the edited video. These descriptions can then be merged with the original database. The new descriptions will show up in the Stratagraph as additional Stratum for the video that was used in the story.

The inefficiency of the annotation process, mentioned at the start of this section, is due in part to the context that a logger is placed in when trying to describe a volume of video. It is a difficult task to imagine all the possible uses for a piece of video, but this is the job of the logger. By using Homer, the problem of context becomes less significant. When a maker is creating a story model using Homer, they are thinking about how to create a story from the available footage in the database. Because the maker is engaged in the process of building story, any descriptions they incorporate into their model should be useful descriptions for building stories by design. While a conventional logger must try to place themselves in the context of building stories, the logger that uses Homer is already operating in that context.

Once a story model has been built, it can be used to annotate the entire database. By adding a Block that excludes all shots that have already been used by this model, the story model becomes an annotation engine. Every time a new edit is rendered using this model, Homer will select new footage. By rendering several edits with this modified story model, the user can quickly apply the descriptions in the model to all the applicable footage in the database. This process also has the beneficial side effect of structuring the video database around the story used for logging. It is this structuring process that enables Homer to be used as an information tool for non-expert users.

Stories as Information

The process of using Homer to annotate with a variety of story models can be viewed as a way of structuring a database around several types of stories. The story models that are used for annotation add their descriptions to the database. If these models are used later to create stories from the same database, they should have a large set of appropriately described video to choose from.

In the travel example, the database would need to be structured by a number of models created by editors or story makers. Then, both the database and the models used to structure it would be supplied to the travel agent. Since the travel agent is not trained as an editor, they may not have enough knowledge of story to design their own models. It would be an easier task for the travel agent to adjust an existing model to the preferences of their client. Therefore, the models used to structure the database are included to give the travel agent a base to develop stories from.

Using Homer to create stories for an end consumer rather than for a maker is a more challenging scenario. When an editor uses Homer, the editor can correct any problems with a story that Homer generates. But if Homer produces a flawed story for a viewer, Homer has essentially failed in the task of relaying information to the viewer.

There are two basic ways in which Homer can falter in the creation of a story. Either the selected content can be wrong or the content can be arranged incorrectly. The process of structuring the database using Homer for annotation should handle the case of inappropriate content. But the problem of sequencing the data is in some ways a more subtle problem that deals with low-level knowledge of editing. The descriptions generated from decisions made using the Sequencer add information to the database that helps Homer to deal with the problem of arranging content.

Conclusion

The video database project described in this paper gives the user powerful control over the video medium. Using the toolset, stories can be pulled from a database of annotated video using abstract story models. These same models can be used to restructure the database around a better understanding of story. The graphical nature of the toolset allows the system to be used as an informational tool by non-expert users for applications such as the aforementioned video travel database.

For all the ideas that this research has produced, it has also raised some questions. The most important of these may be the future of video annotation. This paper discusses one alternative to conventional annotation which this toolset allows. Though the means of entry may have changes, annotation still lies at the heart of video databases. In the system described in this paper, annotations can be added to any group of contiguous frames of video. Although this is an important type of flexibility, video stories still rely heavily on editing. In editing, it is the juxtaposition of shots which is important, not the isolated elements. The incorporation of annotation of these transitions may be the next important step in the evolution of video annotation. When this system is expanded so that any sequence of juxtaposed shots can have annotations attached to it, the user will enjoy another advance in their control over the video medium.

Acknowledgements

This research has been funded by a grant from British Telecom.

References

Goldman-Segall, R. (1993). Interpreting Video Data: Introducing a "Significance Measure" to Layer Descriptions. Journal for Educational Multimedia and Hypermedia.

Haase, K. (1993). Framer: A Persistent Portable Representation Library. In AAAI-93.

Morgenroth, L. (1992). Homer: A Story Model Generator. BSCS, MIT.

Pentland, A., Picard, R., Davenport, G., Welsh, R. (1993). The BT/MIT Project on Advanced Image Tools for Telecommunications: An Overview. In ImageCom 2nd International Conference on Image Communications, Bordeaux, France.

A. Pentland, R.W. Picard, S. Sclaroff (1994). Photobook: Tools for Content-Based Manipulation of Image Databases. In SPIE Conf. Storage and Retrieval of Image and Video Databases II, No. 2185, San Jose, CA.

Smith, T.A. (1992) If You Could See What I Mean… MS Thesis, MIT.