Automatic Digital Video Production Concepts[1]

G. Ahanger and T.D.C. Little

Multimedia Communications Laboratory
Department of Electrical and Computer Engineering
Boston University, 8 Saint Mary's Street
Boston, Massachusetts 02215, USA
(617) 353-9877, (617) 353-6440 fax
{*gulrukh,tdcl*}*@bu.edu*

MCL Technical Report No. 09-02-1998

Abstract–Video production involves conceiving a story, shooting raw video footage, and editing the final piece. Editing involves manually cutting frames and frame sequences from the raw video and composing the sequences with special effects to render the production sequence of the original story concept. The introduction of digital video technology now greatly expedites this process and allows for novel automatic manipulation of recorded video segments without human intervention. One example is the automatic manipulation or composition of news video segments for personalized delivery to individuals.

A digital video production system (DVPS) encompasses the acquisition, storage, selection/editing, composition, and customization of video data. A DVPS cannot operate completely automatically as a human needs to participate in the creation of the video data; however, we believe that it is possible to automate the subsequent steps with suitable restrictions and assumptions on the process. In this chapter, we describe the video production process and the implications for the process due to digital technology. We also describe the requirements of an automatic digital video production system to achieve editing, composition, and customization.

Keywords: Video production, editing, composition, customization, news automation.

[1]To appear in *Computer Information and Technology*, 1998. This work is supported in part by the National Science Foundation under Grant No. IRI-9502702.

1 Introduction

Conventional video production consists of three phases: preproduction, production, and post-production [14]. The preproduction phase involves conceiving a story idea, conducting background research, and laying out the desired pieces. In the production phase raw footage is shot according to the script. Finally, the post-production phase involves editing the footage, dropping shots/frames, and sequencing the shots to achieve desired effect and narrative (story). These three phases are involved in all video production systems including current electronic news gathering (ENG) techniques for television. The ENG process originated in the 1930s using drawings, photographs, and newsreel footage produced for motion picture distributors in production. Later, networks relied on wire services and newsreel companies for their footage. The latency between event occurrence and airing was reduced with the introduction of high quality portable cameras, the means for rapid editing, and live video satellite feeds. ENG continues to be in transition; however, the basic methods of gathering and assembling a video-based news story has remained the same.

The development of next-generation video production systems is heavily influenced by the capabilities of new digital technology. For example, higher network bandwidth, streaming-enabled data transfer protocols, large-scale storage servers, digital video capturing equipment, high compression rates, and high-end multimedia-enabled workstations/terminals all are fueling this change. Once in the digital domain, video scenes/shots can be manipulated through automatic on-line selection, editing, assembly, and dissemination. Video that has been used to create a movie or a news story no longer needs to be used in a single rendering but can be used in multiple contexts without involving an extensive reproduction process.

In conventional production systems, a human decides which video segments should be used and how they should be assembled. In an automatic *digital video production system* (DVPS) these decisions are mechanized. To select and compose video segments, sufficient information must be maintained about each segment to support the composition algorithms. A DVPS requires modeling, capturing, storing, editing, managing, composing, and delivery of video data. Each application, based on its objective, requires a data model and ontology to support such composition. The data model and ontology specify the types of segments that are required by the application, the type of information to be extracted for dynamic searching capability, and the relationships among data required for composition. All of these data are managed by a digital video database or archive.

Newscasting, sportcasting, and distance learning are examples of on-demand applications

that will be enabled with the use of archives of digital video. The World Wide Web can serve as a front end to these applications with streaming delivery of video increasingly viable due to better network bandwidth and video compression techniques that improve the quality of image transmission. At the request of a user, appropriate video segments from multiple video pieces can be selected, composed, and delivered and therefore, many compositions are possible instead of the single composition of traditional video production. The major benefit of this approach is a better mapping of user preferences to delivered content, including the incorporation of advertising materials.

Recent regulatory and industrial developments include the FCC's efforts to convert TV stations from analog to digital [18] and the integration of Web-based technology with the TV. WebTV [6, 18] is an example of such an integration. Enabling technologies such as Microsoft's Broadcast Architecture for Windows claim to allow an individual to choose content from the Internet, local TV, cable TV, and other sources. Satellite-based systems also promise to provide a means for both interactive, Internet-based services and digital TV. These indicators suggest that viable applications based on digital video archives are not far out on the horizon.

There are a number of open issues involved in the shift from a conventional video production to a DVPS. In the following we discuss these issues and the techniques associated with video production, including editing, composition, and customization.

The remainder of this chapter is organized as follows. In Section 2 we describe the processes involved in conventional video production. In Section 3 we describe the steps involved in semiautomatic video production. In Section 4 we consider approaches to evaluate results of automatic composition including performance metrics. Section 5 concludes the chapter.

2 Conventional Video Production Systems

The production of a complete video[2] piece involves a number of phases as illustrated in Figure 1. In the preproduction phase, before creating a video, an underlying concept or storyline is developed that serves as a guide for production efforts. For example, in electronic news gathering, a storyline is developed based on a current event or other cultural, social, political, or experimental curiosity [14].

[2]In this chapter we use the terms "video" and "video data" interchangeably. Also, for simplicity of discussion, we consider audio recorded with video to be an integral video component, even though audio composition (e.g., voice-overs), is an important aspect of video production.

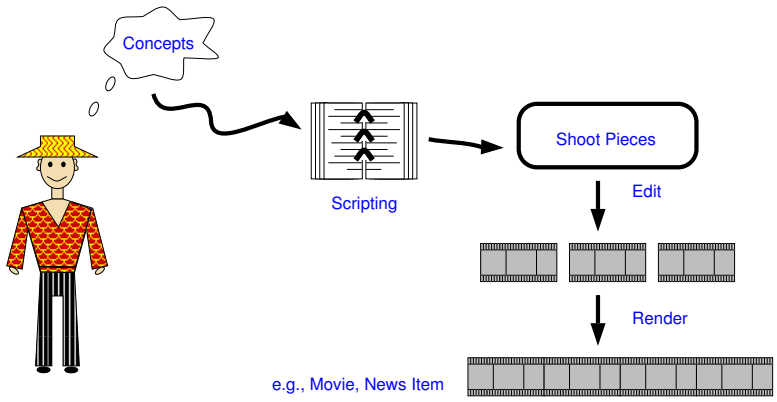

Figure 1: The process of video production.

In the next phase, shots that create a beginning, middle, and an end of a story are formalized conceptually. Once these items are determined, they are scripted. A script contains detailed instructions of how and what is to be shot and serves to minimize effort in the shooting process. For example, in ENG, a script can span many news items and can consist of many pages of text.

The production phase involves the shooting of raw video footage. The location is prepared, equipment is set up, and lights are arranged. A shot is composed while taking care of balance and symmetry. Then the actual film shooting occurs and information about the shot is logged. The process is repeated until all desired footage is complete, including shots recorded to provide continuity between core pieces.

In the post-production phase, the raw video is manipulated and prepared for distribution. In ENG or documentary-making a post-production script is prepared that describes the information gathered and any necessary background information. The script can be written before or during editing. This script is read (e.g., by an anchor person) in conjunction with the edited video. Usually the prerecorded video shots are delivered to an editing point where shots or frames are cut and composed for the final piece.

There are three methods of assembling shots [14]: cut-only, A/B roll, and nonlinear. Cut-only assembly makes an instantaneous transition between two shots. If more subtle transitions are required (e.g., fade, dissolve, and wipe) the A/B roll editing is used. In this technique alternate shots are played from first (A) machine and then from the other (B) machine and fed through a video switcher that creates special effects. Nonlinear editing, in the basic application, requires a multiple-source system and computer control to yield

these and other complex transitions. Once a film or video is edited, it is then rendered to videotape or distributed via satellite to end-users.

3 Semiautomatic Digital Video Production Systems

Out of the three stages involved in video production, only the post-production stage can be automated. The process of actual production (i.e., shooting raw video footage) requires human involvement. Therefore, digital video production is, at best, a semi-automated process. Editing and composition of digital video data are accomplished in the post-production stage. However, before we realize our vision of dynamic and automatic composition of video, we need to resolve a variety of issues surrounding this goal. A semiautomatic system requires information about content for editing and composing a piece of video. The system also requires techniques for composition. Therefore, identifying the information sufficient for editing and composition, determining how the information should be extracted, and creating techniques for cohesive video composition must be addressed.

In a semiautomatic DVPS, a user access scenario is as follows: video segments are selected in response to a user's request and selected segments are composed to create a story that follows a logical sequence including a beginning, middle, and an end. For a dynamic application, the key problem of sequencing segments falls into the post-production scripting domain. The digital video editing and composition is supported by number of computer-based tools. These tools aid a human operator in post-production but do not automate the process. In the following subsections we briefly discuss these tools and describe techniques necessary for full automation of the process.

3.1 Digital Video Editing

Conventional video editing is based on linear access to magnetic tape performed by a professional. Nonlinear editing is achieved by digitizing analog video data to a random-access medium such as magnetic disk. Today many personal-computer-based solutions exist for this purpose. Segments can be easily recorded and manipulated with special effects such as wipes, dissolves, fading in/out, distorting, and embossing. Digital video editing packages such as Adobe Premier, Kohesion, and MediaStudio provide robust tools for commercial and professional use [26]. In addition to functions for selection, transitions, and trimming,

operations including ripple and rolling edits, multiple-track selection, jog, shuttle, and play enable large amounts of footage to be quickly edited.

Some of the commonly available digital video formats include QuickTime, used in Apple's multimedia technology; Motion-JPEG, proposed by the Joint Photographic Experts Group; AVI, used by Microsoft in its multimedia applications; MPEG, defined by the Motion Picture Experts Group and including multiple formats (notable are MPEG-1, 2, and 4); RealVideo, defined by RealNetworks and suitable for low-bandwidth, high-latency environments like the Internet; and VXTreme, another streaming solution by VXTreme Inc. Most editing tools use AVI, M-JPEG, MPEG, or QuickTime formats. Typical of the tools, Avid Cinema [21] combines video editing software with a video input/output card. Analog video data are converted into digital data for editing and then converted back to analog for playout as a NTSC TV signal. Video data are compressed in the M-JPEG format. MediaStudio and MGI VideoWave are some of the first packages to make use of MMX processor to speed up the rendering processes. Recent development in 3D graphics cards also aids in improving rendering speed [12], hence, speeding the editing process.

The aforementioned tools are designed to facilitate editing by a human operator. To support dynamic video composition and delivery we require automatic selection and composition of video segments from an archive. The selection of segments from various sources to form a new piece of video is called repurposing (Figure 2). To repurpose a segment without human intervention, sufficient information must be associated with the raw video data. Therefore, we require tools to extract and annotate information about video segments. To support news video production, where time-to-production is critical, the process of video data annotation must be fast and efficient. The process of information extraction and related issues are discussed next.

3.2 Video Processing for Information Extraction

Video processing for information extraction is a two-step process: establishment of information to be extracted and the extraction process itself. The first step involves creating a model for information to be extracted (Figure 3). Once this is defined, techniques for extracting information consistent with the model are applied. These are described below.

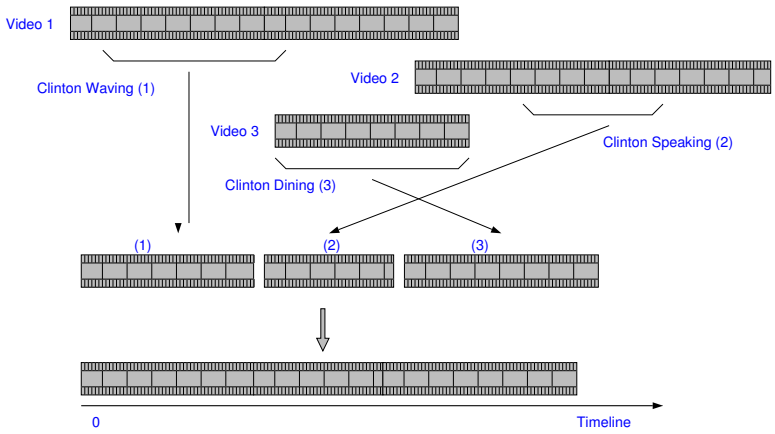

Figure 2: An example of video repurposing.

3.2.1 Data Models and Ontologies

The information that needs to be extracted from video data is dependent on the domain-specific application. For example, in a slide show, we require knowledge of structural information, type of slides, and their order. In a comedy clip, we need information about the properties and behaviors of the objects involved. Therefore, before information extraction, we need to establish the functionality of an application, what information will support the functionality, and how different types of information are related. Next, an object ontology can be defined based on the application domain and the aforementioned functionalities. The defined ontology facilitates queries such as "retrieve a video of political meeting between Clinton and Blair in the White House," beyond simple keyword-based methods. Luke et al. [13] define one such ontology for use by Web agents. An example of their ontology is shown below.

```
<ONTOLOGY ‘‘out-ontology’’ VERSION = ‘‘1.0’’>
<ONTDEF CATEGORY = ‘‘Person’’ ISA = ‘‘org.Thing’’>
<ONTDEF RELATION = ‘‘lastName’’ ARGS = ‘‘Person STRING’’>
<ONTDEF RELATION = ‘‘firstName’’ ARGS = ‘‘Person STRING’’>
<ONTDEF RELATION = ‘‘marriedTo’’ ARGS = ‘‘Person Person’’>
<ONTDEF RELATION = ‘‘employee’’ ARGS = ‘‘org.Organization Person’’>
</ONTOLOGY>
```

The above ontology provides classification of "person" and "organization" and relation-

ships of "marriedTo" and "employee." This ontology can be used to retrieve results for queries like "retrieve documents by Cook who is married to George Cook and works for UMD."

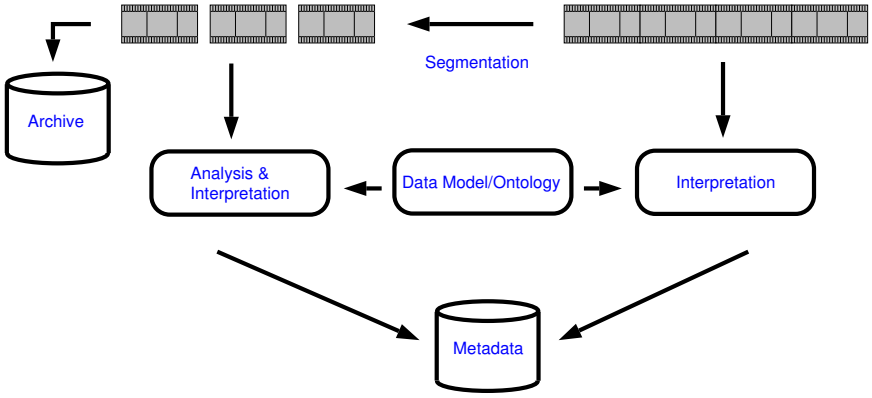

Figure 3: Schematic of video data processing.

In the video domain, we need to form an ontology that possesses sufficient semantics for both information extraction and retrieval. Apart from the ontology, a model for video data also needs to be defined. Two types of data models are commonly used: segmented and stratified. In the segmented data model, narrative is isolated into small units and data are accessed between shot endpoints. Shortcomings of segmentation include the absence of overlapping narrative and the inability to achieve the finest grain of decomposition. Information appearing in only a few frames in a segment is associated with the complete segment. In contrast, stratification achieves overlapping sets of shots by contextual descriptions called *strata*. A stratum (Figure 4) provides access to data over a temporal span rather than through shot endpoints.

3.2.2 Information Extraction

Two main approaches can be used to extract information from available data. First, if data exists as discrete segments, each segment can be parsed for information and the information can be stored as metadata. Second, if data exist as segments of video, the segments can be further decomposed into sub-segments. As shown in Figure 3, information about the complete piece (e.g., if the piece is a movie, then information about the title and producer) is stored and the piece is subsequently segmented by identifying segment start and stop

Figure 4: An example of the stratification technique.

points. The identified segments are parsed and associated with specific information as shown in Figure 5.

The process of video data segmentation and information extraction can be achieved by manual, automatic, or semiautomatic techniques. Usually information is extracted off-line and stored in a more readily usable format for access, (i.e., as video data are time-sequential and large, manual information extraction and annotation are time consuming). Many automatic segmentation techniques exist that focus on different aspects of the problem (e.g., [3, 4, 8, 22]). However, the current performance of these techniques is not satisfactory without human supervision.

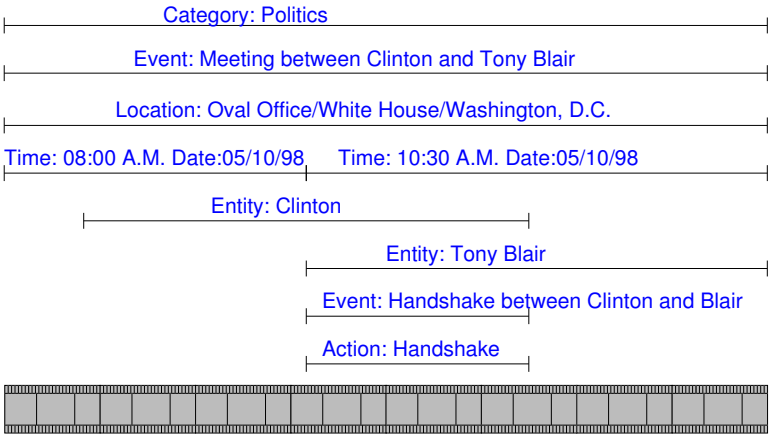

Figure 5: An example of an annotated video segment.

Table 1: Information Extracted for an Object

Object		
ID	=	O61
NAME	=	Saudi Arabia
TYPE	=	Location
METATYPE	=	Country
TIME	=	18:00:57
DATE	=	06/27/96
MEDIUM	=	Video
ORIGIN	=	CNN
POPULARITY	=	0
FILE	=	cnn.mpg
STARTF	=	1000
STOPF	=	3000

Although it originates from human interpretation, closed-captioning of video is another valuable source of information about video content. Closed-captions represent textual transcriptions of the audio track associated with video. Information from closed-captions can be used for automatic selection of video segments for production of a narrative. The closed-caption data stream can be used as follows. Indices of words in each closed-caption stream are maintained in association with the video stream. When a user issues a query to a DVPS, the keywords in the query can be compared against each closed-caption stream. Video segments associated with these matches can be retrieved in response to the query.

One can use a semi-automated tool to support the collection of structured metadata from video sequences. Vane [5] is one such tool that supports both segmented and stratified data models. The tool parses video data files for automatic identification of shots. The parsed files are loaded into an interface that allows the annotator to manually corrects false identifications of segment end points. Two types of metadata are collected: content metadata and structural metadata [10]. Content metadata are concerned with tangible and conceptual entities while structural metadata are concerned with media-specific and cinematographic metadata. In Figure 5, category, event, action, location, and historical time and date are examples of content metadata. In this application, raw metadata are stored in an SGML-compliant format depending on the type of database being used (relational, object-oriented, or semantic). Table 1 shows an example of information extracted for an object and its attributes.

A problem faced during the annotation process is that related content can span multiple

video tapes or video streams, when such limitations exist. A segment can continue in another stream (e.g., in Figure 6, Clinton's speech spans two tapes). Hence, while indexing, one must label the two tapes as containing related content.

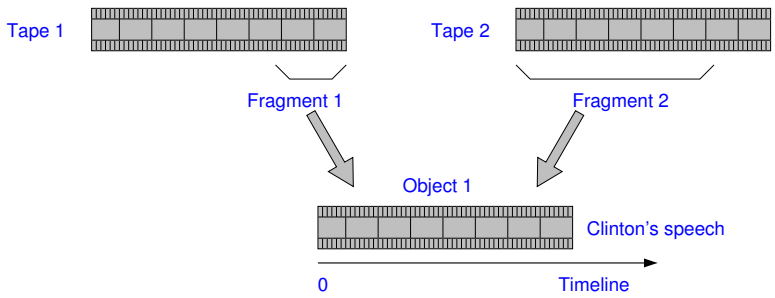

Figure 6: An example of related content spanning two tapes.

This issue can be resolved by treating all of the fragments of a segment as a single object. These fragments are transparent when content is annotated. For example (Figure 6), the concept of Clinton is present in both fragments but we index Clinton once at the time of annotation as Object 1.

Associating time and date with segments proves to be an another major problem, especially in the news video domain. Conceptually, there are two sets of times associated with a segment: one corresponds to when the segment is created/acquired, and the other represents the historical date and time of the segment's event. The two times can be independent of each other because often there is a time lag between an occurrence of an event and the recording of video for the event (e.g., video footage shot after a plane crash). Therefore, a database can contain recently-created segments that have references to a past event. In such cases, for continuous flow of information in a narrative, segments may need to be ordered according to their creation time. Here the creation time should be treated as media-specific metadata and historical time as content-specific metadata.

3.3 Video Composition

Video composition is a process in which selected segments are sequenced to produce a narrative. In conventional video production systems, narratives are manually composed. Applications like movies, newscasting, educational lectures/lessons, and games are produced based on a script. A DVPS must also possess knowledge of how these segments can be

automatically composed to produce a cohesive narrative, i.e., one that has correct semantics and a smooth flow of information. Here we discuss composition techniques from existing production systems.

ConText [7] is a system for automatic temporal composition of a collection of video shots. It lets users semirandomly navigate through a collection of documentary scenes associated with a limited range of content metadata describing *character*, *time*, *location*, and *theme*. The next scene shown to the user is determined based upon a scoring of all available scenes. This scoring aims to obtain the preferred continuity and progression of detail in the presentation. This is made possible by establishing a present context consisting of metadata found in already-played shots or shots chosen by a user. Each metadata entry is associated with a relevance score. The theme, or storyline, is maintained by human intervention and is not completely automated.

ConText demonstrates how cognitive annotations of video material can be used to individualize a viewing session by creating an entirely new version through context-driven concatenation. This dynamic reconstruction can include video material made in a totally different context, thus performing a repurposing of the material.

AUTEUR [15] is an application that is used to automatically generate humorous video sequences from arbitrary video material. The composition is based on the content describing the *characters*, *actions*, *moods*, and *locations*; and, in addition, the information about the position of camera with respect to a character, such as, close-up, medium, and long range shots. Content-based rules are used to compose shots. Hence, a single shot can be repurposed to form different jokes.

Canvass [2] is a news video composition and delivery system. In this system, content-based metadata and structure-based metadata are used to compose a news item. The composition is based on knowledge about the structure of a news item and how various types of segments fit into the structure. Structure is based on an introduction, body, and an end in the narrative. Different segment types can be identified and the composition is determined under the assumption that each class of segment presents information from a different viewpoint (e.g., field shots vs. interviews). The segments are defined for independence in playout. That is, related segments can be included or excluded to meet preference to time constraints without sacrificing continuity. Within the restrictions imposed by the composition grammar [1], segments belonging to the body of a narrative can be presented in any order if their creation times are within a small range.

In the following, we describe techniques employed in the aforementioned systems and additional techniques that can be used in a composition system.

Visual rank-based: Rank-based composition is achieved based on a weight assigned to a keyword [7]. Similarity is determined by comparing the keywords or terms representing one video segment with another. If the similarity is sufficient then there is a logical flow of concepts between the two segments and they can be catenated for playout. Consider the example shown in Figure 7; it consists of four video segments represented by weighted keywords.

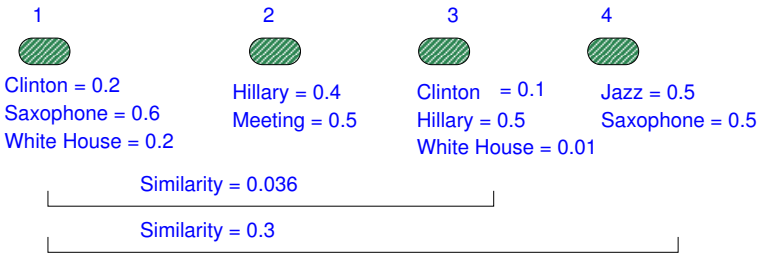

Figure 7: An example of video segment sequencing using a term-weight comparison.

The following keyword vectors are generated by the four segments:

	$Clinton$	$Saxophone$	$Hillary$	$WhiteHouse$	$Jazz$	$Meeting$
Seg_1	0.2	0.6	0	0.2	0	0
Seg_2	0	0	0.4	0	0	0.5
Seg_3	0.1	0	0.5	0.01	0	0
Seg_4	0	0.5	0	0	0.5	0

Salton's cosine metric is used to sequence the segments for playout [23]. Each segment is represented by vector ss_i of contained terms or words. The vector q_j generated from the keywords within a query is used for comparison. The metric uses the cosine of the angle between the two vectors in the multidimesional term space of t.

$$\text{cosine}(ss_i, q_j) = \frac{\sum_{k=1}^{t}(term_{ik} \times term_{jk})}{\sqrt{\sum_{k=1}^{t}(term_{ik})^2 \times (term_{jk})^2}}$$

$\text{cosine}(Seg_1, Seg_2) = 0$

$\text{cosine}(Seg_1, Seg_3) = \frac{0.2\times0.1+0.2\times0.01}{\sqrt{(0.2+0.6+0.2)^2\times(0.1+0.5+0.01)^2}} = 0.018$

$$\text{cosine}(Seg_1, Seg_4) = \frac{0.6 \times 0.5}{\sqrt{(0.2+0.6+0.2)^2 \times (0.5+0.5)^2}} = 0.21$$

Segment 1 is most similar to segment 4; hence, segment 4 is sequenced following segment 1 for playout. User feedback can also be used to sequence the segments for playout. For example, in Figure 8, a user might change the weights assigned to some of the keywords representing segment 1. Using the cosine similarity again, based on the new weights, segment 3 is sequenced to be played out after segment 1.

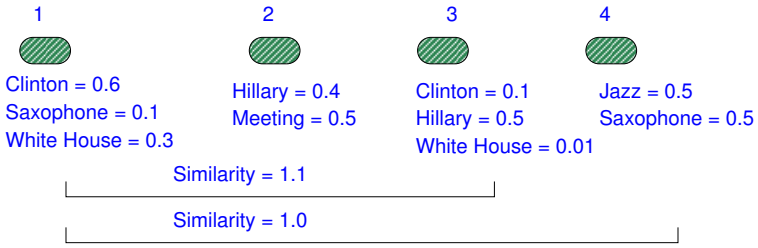

Figure 8: An example of video segment sequencing using user feedback.

Text rank-based: In addition to ranking annotations from visual content, information contained in audio can be ranked. Audio can be converted to text with the help of speech recognition tools [20] or closed-captioning. Using text indexing techniques, the indices can be formed for similarity based retrieval. Each term or word is ranked based on the frequency of its occurrence (TF_{ik}) in a document. Depending on a domain, the ranking can be made sensitive to the document length or the complete population of documents (DF_k) in the database in which the term appears. The following metric, sensitive to the document population, is most commonly used:

$$Rank_{ik} = \frac{TF_{ik}}{DF_k}$$

This metric finds the rank of the k_{th} word in the i_{th} document.

For the Canvass application text ranked-based retrieval is used in addition to the visual ranked-based retrieval (rank is 0 or 1). Depending upon keywords used in a search, closed-captioned documents are retrieved based on the similarity with the query. A threshold is established, and documents with similarity values below the threshold are not considered. A set of keywords can often represent multiple news events. For example, multiple events

associated with the keyword "Clinton" are present. Therefore, the retrieved documents are clustered with secondary levels of similarity. Video data associated with these documents are then used in the composition.

Temporal-based: Temporal information can also be used to sequence segments for composition. Consider the query, "Retrieve segments with Clinton waving to the crowd after he presents a speech." The following options exist for playout of the desired information. First, a clip with this scenario (Figure 9) exists in the available data set. Second, a segment containing Clinton waving and another segment containing Clinton giving a speech are separately available. In the second case, both of these segments can be selected and sequenced in the desired order.

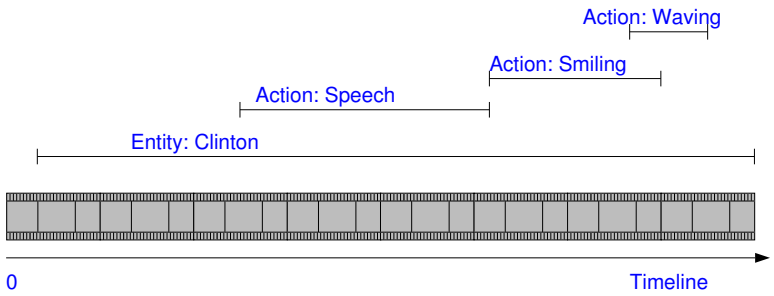

Figure 9: An example of video segment sequencing using temporal ordering.

Rule-based: In the rule-based scheme, additional restrictions are imposed on the sequencing of content. Not only are the desired segments presented, but the segments are sequenced according to the rules imposed on compositions. We divide the rules into two types: content-based and structured-based.

Content-based: The composition is achieved by imposing rules on content or information contained in video clips. For example, proof of a theorem cannot be presented before the problem statement. Nack et al. [15] impose content-based rules on the automatic joke composition. Ozsoyoglu et al. [19] impose content-based rules to drop or include the segments in a composition.

Structure-based: Structure-based rules describe a composition based on a general architecture of a narrative (e.g., an introduction, field shots, commentary, and a conclusion).

A shortcoming of the content-based rules approach is requirement for a set of rules for each and every scenario. The number of rules increases with the number of scenarios. It is not reasonable to establish rules for incoming data, especially when the time between acquisition and delivery is short. Structure-based rules overcome this constraint with a finite set of rules.

The above limitations can be overcome by imposing rules based both on the content and structure of an application domain. For example, in the Canvass application we identify the structure of a news item and, based on this structure, we divide news segments into different types. Based on content, segments retrieved as a result of a query are clustered, and, based on the structure, segments within each cluster are sequenced for playout. The advantage of this technique is that we require only a single set of rules that do not change.

Playout duration can also be limited. To adjust the playout temporal constraints, it is possible to use the type of segment and its importance to the structure to achieve the composition. To maintain thematic continuity, we use a segment's creation time and its measure of similarity among the segments for sequencing. We quantify thematic continuity as a smooth progression of a storyline while maintaining temporal continuity.

Next, we present rules that can be applied to a news video database for a news item composition. The segment types identified in the news structure are shown in Table 2.

Table 2: Structure of a News Item

<table>
<tr><td colspan="3">Headline</td></tr>
<tr><td colspan="3">Introduction</td></tr>
<tr><td rowspan="5">Current (body)</td><td colspan="2">Comment</td></tr>
<tr><td colspan="2">Wild-scene</td></tr>
<tr><td>Interview</td><td>Question&Answer (QA)</td></tr>
<tr><td colspan="2">Speech</td></tr>
<tr><td colspan="2">Enactment</td></tr>
<tr><td colspan="3">Enclose</td></tr>
</table>

Let s_h be a segment representing a headline, s_i be a segment representing an introduction, s_b be a segment representing a body, and s_e be a segment representing an enclose. The following rules define composition.

1. $\{s_h\} = NI$: Only a headline can be present in a news item NI.
2. $\{s_h, s_i\} = NI$: Only a headline and an introduction can be present in a news item NI.

3. $\{s_i\} = NI$: Only an introduction can be present in a news item NI.

4. $\{s_h, s_i, \{s_b^1, s_b^2, ..., s_b^n\}\} = NI$: Only a headline, an introduction, and segments belonging to the body can be present in a news item NI.

5. $\{s_i, \{s_b^1, s_b^2, ..., s_b^n\}\} = NI$: Only a headline and segments belonging to the body can be present.

6. $\{s_h, s_i, \{s_b^1, s_b^2, ..., s_b^n\}, s_e\} = NI$: All segment types are present.

7. $(\forall_i : 1 \leq i \leq n : s_b^i \notin NI) \Rightarrow (s_e \notin NI)$: If a body is not present, then we do not include a segment belonging to Enclose.

Composition rules can also be incorporated in a script. Scripting can be done at various levels depending on the end application. Scripts can be written for on-site shooting (e.g., movies and documentaries). Post-production scripting can include scripts for dynamic composition of available digital video data or scripts for interaction among objects (images, text, video, etc.). For example, scripts are used in Macromedia Director [25] to specify object behavior and properties. The scripts can be easily written by visual operation (e.g., drag and drop and drawing paths for object movements). Scripts can also include operations like activating a pause as a particular frame is being displayed. TV Markup Language (TVML) is a scripting language for use in creating TV programs at NHK [9]. For such scripts, contents of the program are represented by text-based commands such as "display title no. 1" or "zoom in," etc. A computer interprets a script line-by-line to generate TV programs in real-time using multimedia techniques such as real-time computer graphics, voice synthesis, and video playout. Rivl [27] is another language that can be used in a similar manner, in this case to insert special effects by computing composited images in a segment and then sequencing the segments for playout. For automatic composition, Canvass uses a grammar that creates a script based on production rules for creating a newscast.

3.4 Video Customization

In dynamic assembly of content, it is possible to adapt the retrieved information to an individual's specification and a system's capabilities. Information for customization can be acquired either by implicit or explicit techniques. For explicit techniques [11] a user profile is directly acquired from the user. For implicit techniques [17, 24, 28] a user profile is acquired

by observing the behavior of the user (e.g., information about a user's content preference and the order of the presentation). A user profile is used to achieve customization. Content-based and temporal-based customization techniques have been commonly used [11, 17, 24, 28]. In Canvass we have implemented a third type of customization called structure-based customization.

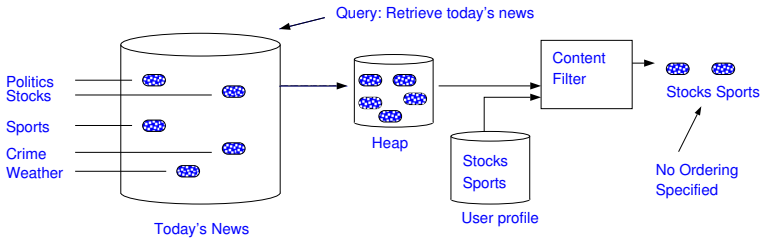

Figure 10: An example of content-based customization.

Content-based customization: According to a user's preference for a certain type of information, only the preferred information is retrieved and all other information is filtered out. Content matching the user's preference is illustrated in Figure 10.

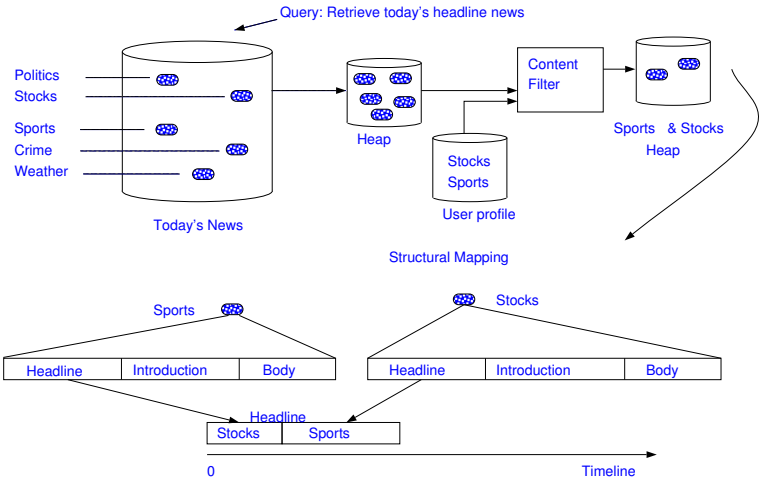

Figure 11: An example of structure-based customization.

Structure-based customization: We define this type of customization as filtering based on structural unit type (e.g., field shots). Figure 11 illustrates an example of a structure-based customization.

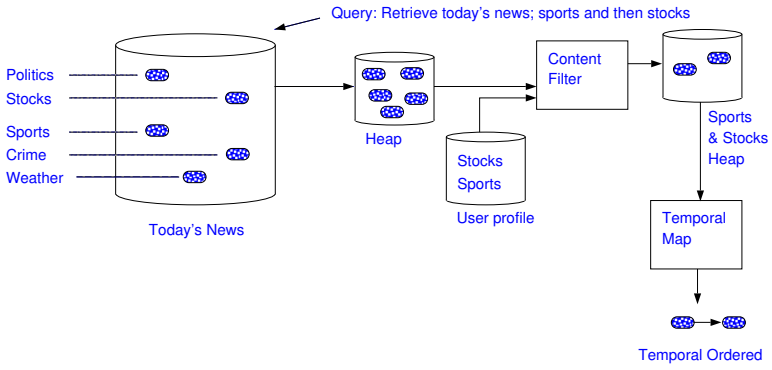

Figure 12: An example of time-based customization.

Time-based customization: There are two types of time-based customization: customization based on playout temporal order and customization based on playout temporal duration.

Temporal order: Customization is achieved by specification of the *relative position* of segments on a timeline as shown in Figure 12.

Time duration: Customization is achieved by specification of playout duration (e.g., the query "recap today's news for two minutes"). If the playout duration of the available data is more than the requested duration, some data need to be dropped [19]. Figure 13 illustrates how data are dropped iteratively to achieve the target duration if there are no interdependencies in the presentation.

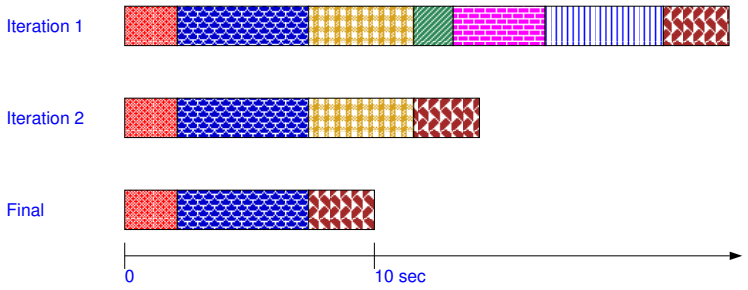

Figure 13: Iterations to achieve a target playout duration.

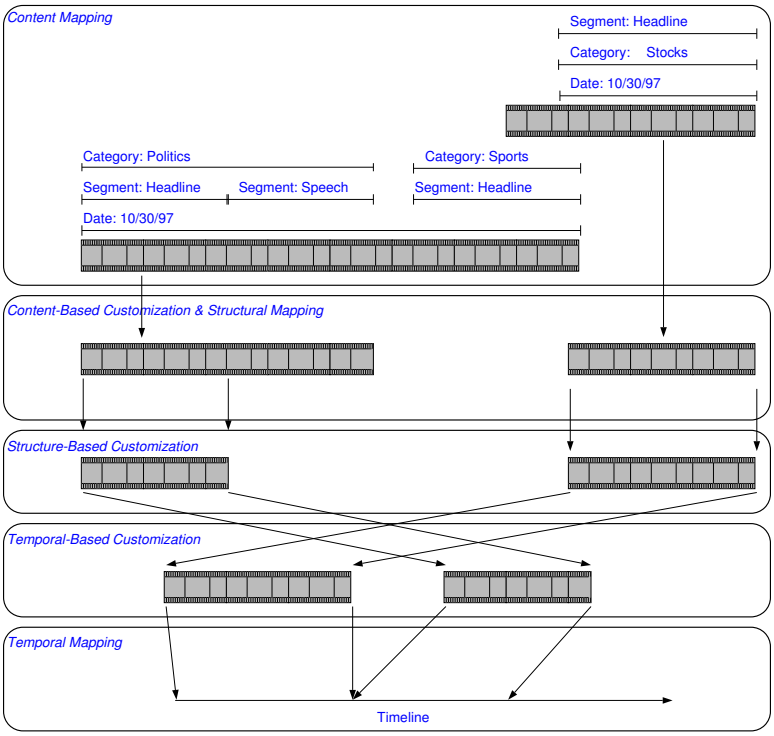

Figure 14: A schematic of newscast composition and customization.

In the following, we recapitulate the overall process of automatic newscast composition and customization. The following query is used for this discussion.

Query: Present the latest news about stocks and politics for a duration of two minutes.

Let the latest news be from 30 October 1997. To compose the newscast, all data corresponding to this date are retrieved. Next, the content that does not satisfy the query specification is filtered out. Figure 14 shows content belonging to politics, stocks, and sports. Therefore, the content belonging to sports is filtered out during content-based customization. Next, the remaining content is mapped to the respective segments. The system analyzes the types of structural segments and the playout duration of each segment. Structural segments that form cohesive information in less than two minutes are retained and the rest are dropped. This process is called structure-based customization. The structures are ordered according to the temporal preference in temporal-based customization. Finally, the structures are mapped to a timeline for playout.

4 Metrics

This section is concerned with the evaluation of retrieval and composition efficiency and efficacy of a DVPS. We require means to evaluate how well a DVPS performs as compared to conventional video production systems. To evaluate a system we first understand the objectives of a system and the various system components. Subsequently we focus on the measures that best reflect the system performance.

A DVPS is not only required to retrieve data but also to achieve composition. Therefore, in addition to metrics for evaluation of information retrieval (IR), new metrics are required to quantify video composition. The metrics used for measuring *recall* and *precision* [23] remain valid for IR; however, these metrics are oriented toward boolean evaluation (i.e., a retrieved object either matches a query or it does not). Recall measures the ability of the system to retrieve all relevant data. Precision measures the ability of the system to present relevant data.

Ranked evaluation metrics can also be used to measure retrieval performance. A retrieved object may not exactly match the query but can have a degree of similarity. A rank-based metric can be applied to evaluate multimedia data retrieval. For example, if an image is retrieved, the degree of similarity between the query and the retrieved image can be measured and ranked accordingly. Narasimhalu et al. [16] have proposed metrics for retrieval of multimedia objects. They propose metrics that measure the *rank, order, spread,* and *displacement* of retrieved objects. These are summarized below.

Order: Order quantifies the ability to sequence data items in the retrieved set. In the example below, the system retrieves data in an incorrect order.

Example 1.
Correct response: $o_1, o_2, o_3, o_4, ...$
Actual response: $o_2, o_4, o_3, o_1, ...$

Rank: Rank measures the degree of relevancy of the retrieved set to the query. In the example, below the rank of individual objects in the retrieved set is less than the actual rank.

Example 2.
Correct response: $o_1, o_2, o_3, o_4, ...$
Actual response: $o_7, o_2, o_4, o_3, o_1, ...$

Spread: Spread measures the shift in the position of a data object in the retrieved set as compared to the correct position as illustrated in Example 3.

Example 3.
Correct response: $o_1, o_6, o_2, o_3, o_4, ...$
Actual response: $o_1, o_2, o_8, o_9, o_3, o_4, ...$

Displacement: Displacement measures the position of a data object in the retrieved set as compared to its correct position as illustrated in by Example 4.

Example 4.
Correct response: $o_1, o_2, o_3, o_4, ...$
Actual response: $o_1, o_2, o_4, o_3, ...$

Some of the above metrics are redundant for the evaluation of a DVPS. After the degree of similarity has been established in retrieval, the ordering becomes trivial and segments are reordered to create a narrative. Spread and displacement metrics are another means of specifying the performance of recall and ranking of the system. Therefore, spread and displacement provide little added information about the performance of the system and can be discarded as useful metrics in this context.

Additional metrics are required to quantify composition performance. A storyline or a theme must be maintained in a composed piece. Hence, a metric that can measure the thematic continuity of the composition must be established. Table 3 summarizes additional metrics for evaluation of composition.

The above metrics are used to evaluate the quality of presentation of an automatic composition. These metrics are especially important when the retrieved data (candidate set) cannot be incorporated in a composition and the results must be culled.

5 Summary

The phases involved in a semiautomatic digital video production are the same as for conventional video production. However, in a conventional system, a human makes all decisions about selecting video segments, composition, and customization. In a DVPS, segments matching a user's query are selected by comparing associated metadata. Start and stop offsets are established from these metadata. Using the offsets, the segments can be extracted

Table 3: Proposed Metrics

Metric	**Explanation**
Information	Measures the amount of information contained in all the segments in a composition with respect to information contained in all the candidate segments not included in the composition.
Playout duration	Measures the performance of a system in achieving the specified playout duration.
Temporal continuity	Measures the continuity in the chronological time frame of presentation. If large forward jumps exist or any backward jump in time exists, then it is considered a disparity.
Thematic continuity	Measures the progression of a storyline or a theme in a composition.
Content progression	Measures the rate of content change within a composition.
Period span coverage	Measures the performance of the system in covering information from the complete period of the available data.
Structural continuity	Measures the structural integrity of a composition.

and algorithmically sequenced for playout. If needed, special effects can be added between segments during rendering.

Sufficient knowledge and judgment that a human requires to produce a piece of video must be mechanized for automatic production. Therefore, to mechanize a video composition and customization much more functionality is required than available in existing video editing tools. Instead of assisting in production the tools need to make decisions about how a composition should be achieved. To make the tools more effective, information (visual, textual, temporal, and structural) within video must be extracted and provided to the composition system. Based on this information a DVPS must possess knowledge of what segments to retrieve and how to compose the segments. All three types of information, content, temporal, and structural may be required for a coherent composition. In addition to sequencing segments based on the information within the clips, creation times, domain-specific structure, and algorithms to customize information under playout constraints are needed. Finally, a means to judge the quality of compositions is required, for which we propose new metrics apart from what are currently used to evaluate conventional information retrieval systems.

A variety of efforts are underway towards the goal of automatic assembly of video-based content. We expect such systems to be a reality in the near future with significant impact on the speed and quality of video compositions.

References

[1] G. Ahanger and T.D.C. Little, "A Language for the Composition of a Newscast," *Computing and Information Technology*, September 1998.

[2] G. Ahanger and T.D.C. Little, "A System for Customized News Delivery from Video Archives," *Proc. Intl. Conf. on Multimedia Computing and Systems*, Ottawa, Canada, June 1997, pp. 526-533.

[3] G. Ahanger and T.D.C. Little. "A Survey of Technologies for Parsing and Indexing Digital Video," *Visual Communication and Image Representation*, Vol. 7, No. 1, March 1996, pp. 28-43.

[4] A. Akutsu and Y. Tonomura, "Video Tomography: An Efficient Method for Camera Work Extraction and Motion Analysis," *Proc. ACM Multimedia'94*, San Francisco CA, October 1994, pp. 349-356.

[5] M. Carrer, L. Ligresti, G. Ahanger, and T.D.C. Little, "An Annotation Engine for Supporting Video Database Population," *Multimedia Tools and Applications*, Vol. 5, No. 3, November 1997, pp. 233-258.

[6] T. Clark, "WebTV Eyes Fall Launch," *Interactive Week*, June 13, 1996.

[7] G. Davenport and M. Murtaugh, "ConText Towards the Evolving Documentary," *Proc. ACM Multimedia '95*, San Francisco CA, November 1995, pp. 377-389.

[8] N. Dimitrova and F. Golshani, "Rx for Semantic Video Database Retrieval," *Proc. ACM Multimedia'94*, San Francisco CA, October 1994, pp. 219-226.

[9] M. Hayashi and H. Sumiyoshi, "TVML and Hierarchical Structured Method for Program Production," *http://www.strl.nhk.or.jp/publica/rd/rd46.html*, NHK Science and Technical Research Laboratories, Tokoyo, Japan, 1997.

[10] W. Klippgen, T.D.C. Little, G. Ahanger, and D. Venkatesh, "The Use of Metadata for the Rendering of Personalized Video Delivery," in *Managing Multimedia Data: Using Metadata to Integrate and Apply Digital Data*, Amit Sheth and Wolfgang Klas (Eds.), McGraw Hill, 1998.

[11] A. Kobsa and W. Pohl, "The User Modeling Shell System BGP-MS," *User Modeling and User-Adapted Interaction*, Vol. 4, No. 2, 1995, pp. 59-106.

[12] S. Leemon, "The Cutting Edge: Video-Editing Software," *Computer Shopper*, May 1997.

[13] S. Luke, L. Spector, D. Rager, and J. Hendler, "Ontology-Based Web Agents," *Proc. 1st Intl. Conf. on Autonomous Agents*, Marina Del Rey, CA, 1997.

[14] R.B. Musburger, *Electronic News Gathering*, Focal Press, Boston, 1991.

[15] F. Nack and A. Parkes, "The Application of Video Semantics and Theme Representation in Automated Video Editing," *Multimedia Tools and Applications*, Vol. 4, No. 1, January 1997, pp. 57-83.

[16] A.D. Narasimhalu, M.S. Kankanhalli, and J. Wu, "Benchmarking Multimedia Databases," *Multimedia Tools and Applications*, Vol. 4, No. 3, May 1997, pp. 333-356.

[17] J. Orwant, "Heterogeneous Learning in the Doppelgänger user Modeling System," *User Modeling and User-Adapted Interaction*, 1995.

[18] J. Ozer, "Industry Trends: Going Digital," *PC Magazine*, April 21, 1997.

[19] G. Ozsoyoglu, V. Hakkoymaz, and J. Kraft, "Automating the Assembly of Presentation for Multimedia Data," *IEEE Intl. Conf. on Data Engineering*, February 1996, pp. 593-601.

[20] N. Randall, "Computer, Take a Memo," *PC Magazine*, January 1998.

[21] J. Rizzo, "Making Movies," *Creative Mac*, April 1, 1997.

[22] H. Rowley, S. Baluja, and K. Kanade, "Neural Network-Based Face Detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 1, January 1998, pp. 23-38.

[23] G. Salton and M.J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill Book Company, New York, 1983.

[24] B.D. Sheth, "A Learning Approach to Personalized Information Filtering," *M.S. Thesis*, MIT, Cambridge, MA, February 1994.

[25] L. Simone, "The Scripting Standard," *PC Magazine*, November 4, 1997.

[26] L. Simone, "Video Editing," *PC Magazine*, October 7, 1997.

[27] J. Swartz and B.C. Smith, "A Resolution Independent Video Language," *Proc. ACM Multimedia '95*, San Francisco CA, November 1995, pp. 179-188.

[28] T.W. Yan, H. Jacobsen, H. Garcia-Molina, and U. Dayal, "From User Access Patterns to Dynamic Hypertext Linking," *Proc. 5th Intl. World Wide Web Conf.*, Paris, France, May 1996.