Information Visualization within a Digital Video Library

MICHAEL CHRISTEL AND DAVID MARTIN

Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Abstract. The Informedia Digital Video Library contains over a thousand hours of video, consuming over a terabyte of disk space. This paper summarizes the multimedia abstractions used to represent this video in prior systems and introduces the visualization techniques employed to browse and navigate multiple video documents at once.

Keywords: digital video library, information visualization, multimedia abstraction

1. Introduction

The Informedia Project at Carnegie Mellon University deals primarily with video. The goal of the project is to enable full content search and retrieval from digital video libraries (Christel et al., 1996; Wactlar et al., 1997). Consider the task of trying to find a five-minute video clip of interest from a library of a thousand hour-long videotapes. In the analog domain, this task would be interminable and the frustrated user would probably walk away without completing the task. Simply digitizing the video will not make the job easier. Through the use of speech recognition, image processing, and natural language processing, the digital video can be:

- segmented into smaller pieces; each segment consists of a contiguous range of video and/or audio (or an extent of text such as a paragraph or chapter) that is deemed conceptually similar
- analyzed to derive additional data (metadata) for creating alternate representations of the video
- augmented with indices for fast searching and retrieval of segments

This process is illustrated in Figure 1.

The Informedia Digital Video Library System (IDVLS) accesses over a terabyte of video (over 1000 hours playing time), including documentaries from WQED Pittsburgh, instructional videos from the United Kingdom’s Open University, documentaries from various United States government sources, and news from CNN. IDVLS has testbeds in a local Pittsburgh K-12 school and at government offices in Washington, D.C. The user experiences for IDVLS have been reported elsewhere (Hauptmann et al., 1995a; Christel and Pendyala, 1996).

The initial interfaces respected the Visual Information Seeking Mantra of Ben Shneiderman (1996): “Overview first, zoom and filter, then details-on-demand.” IDVLS provided a number of multimedia abstractions for each video segment that communicated information about the segment yet required less viewing time and storage space (Christel et al., 1997b). Users issued a query, overviewed a set of results using these abstractions, “zoomed” into more detailed abstractions, made use of the abstraction to filter down to the desired video sequence, and then saw the details by playing that sequence. The dominant metaphor was the query metaphor: issue a query, and trust that the search engine has high enough precision and recall to return the desired story in a window of the top results.

This paper begins by reporting on the query interface and how multimedia abstractions can help the user deal with less than perfect precision: if the search engine returns a set of results containing both desired and undesired segments, the multimedia abstractions help the user focus in on the desired segments more quickly. The concluding section then discusses how other visualization techniques can supplement multimedia abstractions to provide:

- better browsing and navigation
- examination of a much larger neighborhood in the results space
- video *exploration* as opposed to just library *querying*

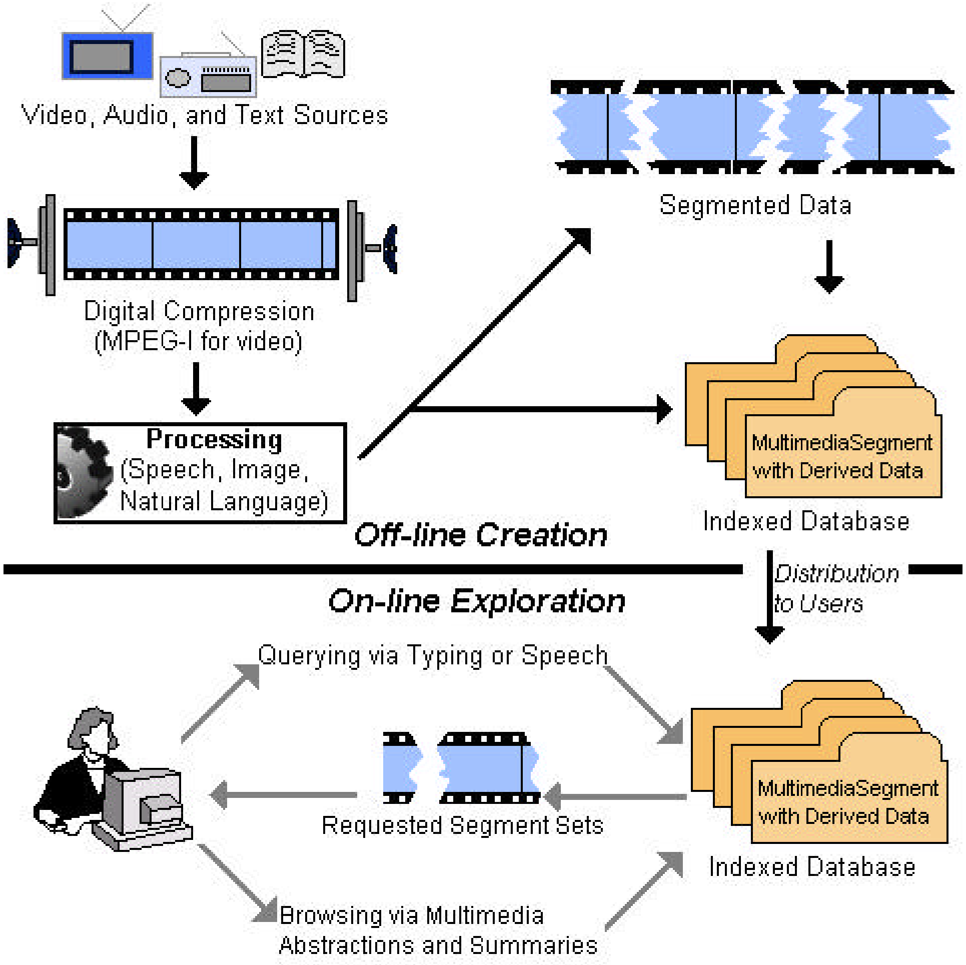


Figure 1. Informedia Digital Video Library System overview

2. Query interface

A wealth of prior information retrieval research, on-line library catalogs, and the World Wide Web search engines have all shown the utility of querying for seeking information within a large corpus. A traditional access method into digital collections has been to issue a query and examine the results. While this method should not be the sole interface into the collection (Furnas, 1996), it was the first interface developed extensively for IDVLS.

The general model for accessing the Informedia digital video library materials through 1997 has consisted of the following steps which may be iterated through many times in a given library session:

1. User defines a problem and selects a corpus for examination.
2. User formulates and submits a query.
3. User examines a single page of up to 30 results. Multimedia abstractions are used to allow the user to make informed decisions more quickly as to which results should be examined in more detail. Each result is a video segment, and visualizations into that segment are provided in the form of headlines, thumbnail images (also called poster frames), series of thumbnail images ("filmstrips"), and skims.
4. Through the abstractions the user selects locations within segments for further inspection.

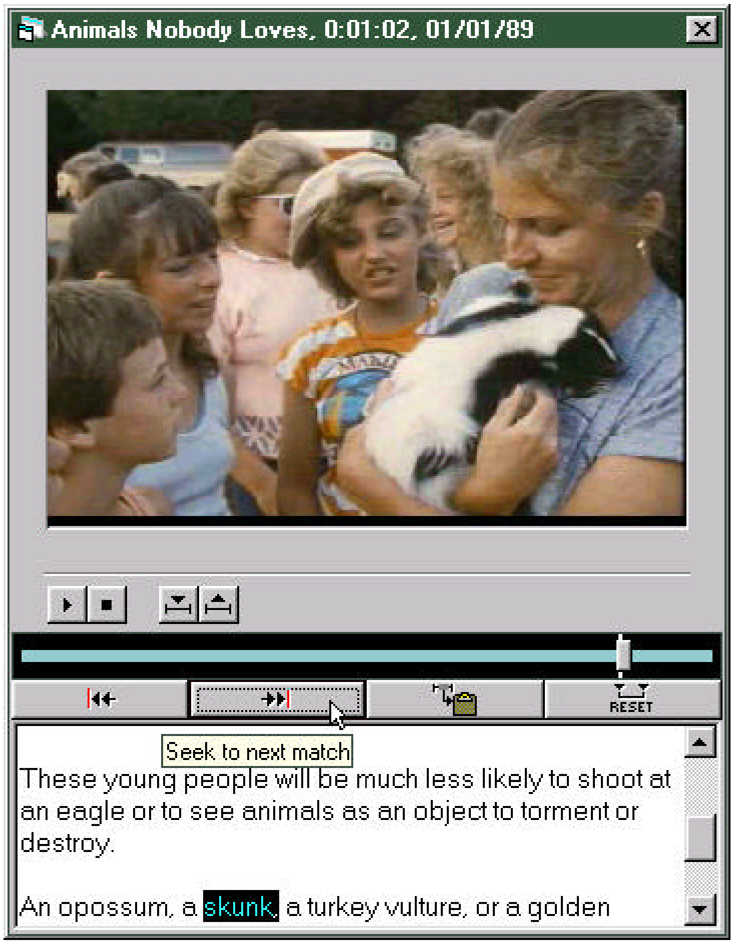
These steps align well with the general model of information seeking outlined by Marchionini (1995).

2.1. Video metadata

During library creation, data is generated to supplement a MPEG-I video (352 X 240 resolution, 30 frames per second). For some video, closed-captioned text is available documenting the narrative for a given video. For others, the Carnegie Mellon University Sphinx-II speech recognizer is used to generate the narrative's text transcript (Hauptmann et al., 1995b). In all cases, Sphinx-II is run against the text transcript to tightly align words to the video. This alignment enables a user to find the precise target of a search. For example, Figure 2 shows how a search on "skunk" can take the user to the location within a segment where "skunk" is mentioned. The transcript is shown beneath the video, and scrolls with the video as the video plays. This figure shows the mouse positioned over the "Seek to next match" button and the tool tips text box describing the action of that button. A user performing a search can play the whole segment for context, or can jump directly to the match points through the use of the buttons in this window.

Image processing is used to decompose a video segment into a set of shots, where *shots* are sets of "contiguous frames representing a continuous action in time or space" (Zhang et al., 1995a). Informedia researchers have developed an algorithm that identifies, for each camera shot, an individual frame within the video that best represents the whole shot (Smith and Kanade, 1996). This frame selection is not tied to the MPEG encoding scheme, as is done elsewhere (Zhang et al., 1995b), and so the term *shot frame* rather than "key frame" is used to signify this representative frame for the shot.

By default the algorithm chooses the shot's middle frame. If camera motion is detected and that motion stops within the shot, then the frame where the camera motion ends is selected. Other image processing techniques, such as those that detect and avoid low-intensity images, further refine the selection process. For the Informedia video library, a video segment consists of a sequence of shots and can be represented compactly by a sequence of shot frames.


Animals Nobody Loves, 0:01:02, 01/01/89

These young people will be much less likely to shoot at an eagle or to see animals as an object to torment or destroy.

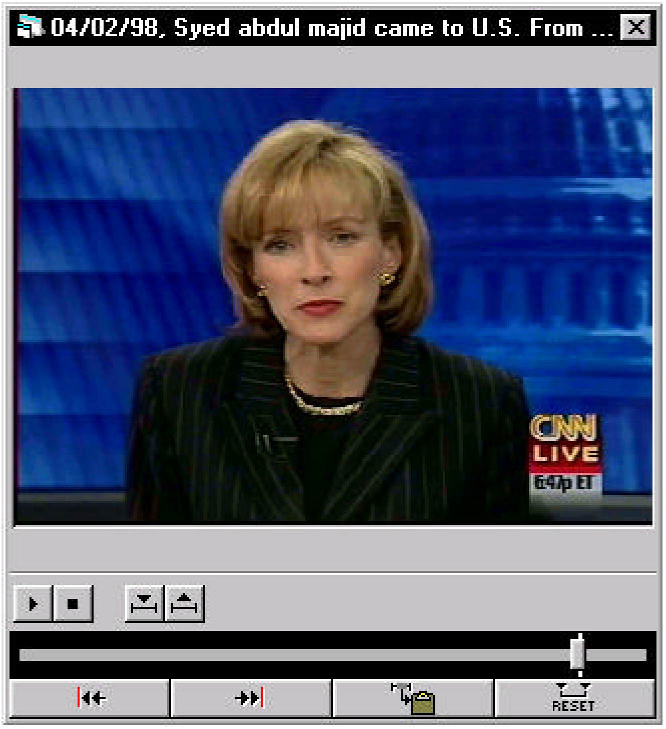
An opossum, a skunk, a turkey vulture, or a golden

Figure 2. Effects of seeking directly to match point, courtesy of tight transcript to video alignment provided by speech recognizer

Image processing also provides additional metadata that documents the video. A region extractor is run across all shot frames to identify the composite objects in the image and their color, shape, texture, and size characteristics (Gong, 1998). Specialized filters are run to automatically detect the presence of faces (Rowley et al., 1995) and text (Sato et al., 1998) in the video. Analogous to the manner in which transcript words are tightly aligned to the video through speech processing, these image processing artifacts are also tightly aligned to the video. The user can search on image features, e.g., a particular anchorwoman's face, and seek to that matching location, as shown in Figure 3.

Natural language processing is used to improve query processing, to generate headlines, and in combination with image and speech processing to perform segmentation on the video. Synonyms and word stemming are available to increase the target set of words for queries, e.g., with the appropriate synonym table and with word suffixing enabled the query "think" will also match " meditate" and "thinking." Headline generation uses the same relevance measure, term frequency inverse document frequency (tf*Idf) (Salton, 1989), as used by the word search engine, but uses this statistical measure to score phrases and then collect the highest scoring phrases into a headline. Video segment boundaries can be determined by changes in image characteristics and changes in topic (Hauptmann, 1998).

Additional metadata includes characteristics of the digitally encoded video, such as the frame rate, duration, and encoding date; and also information entered into a relational database when a video is being digitized, including copyright date, producer, and series title, e.g., “CNN World View” or “The Space Age”.



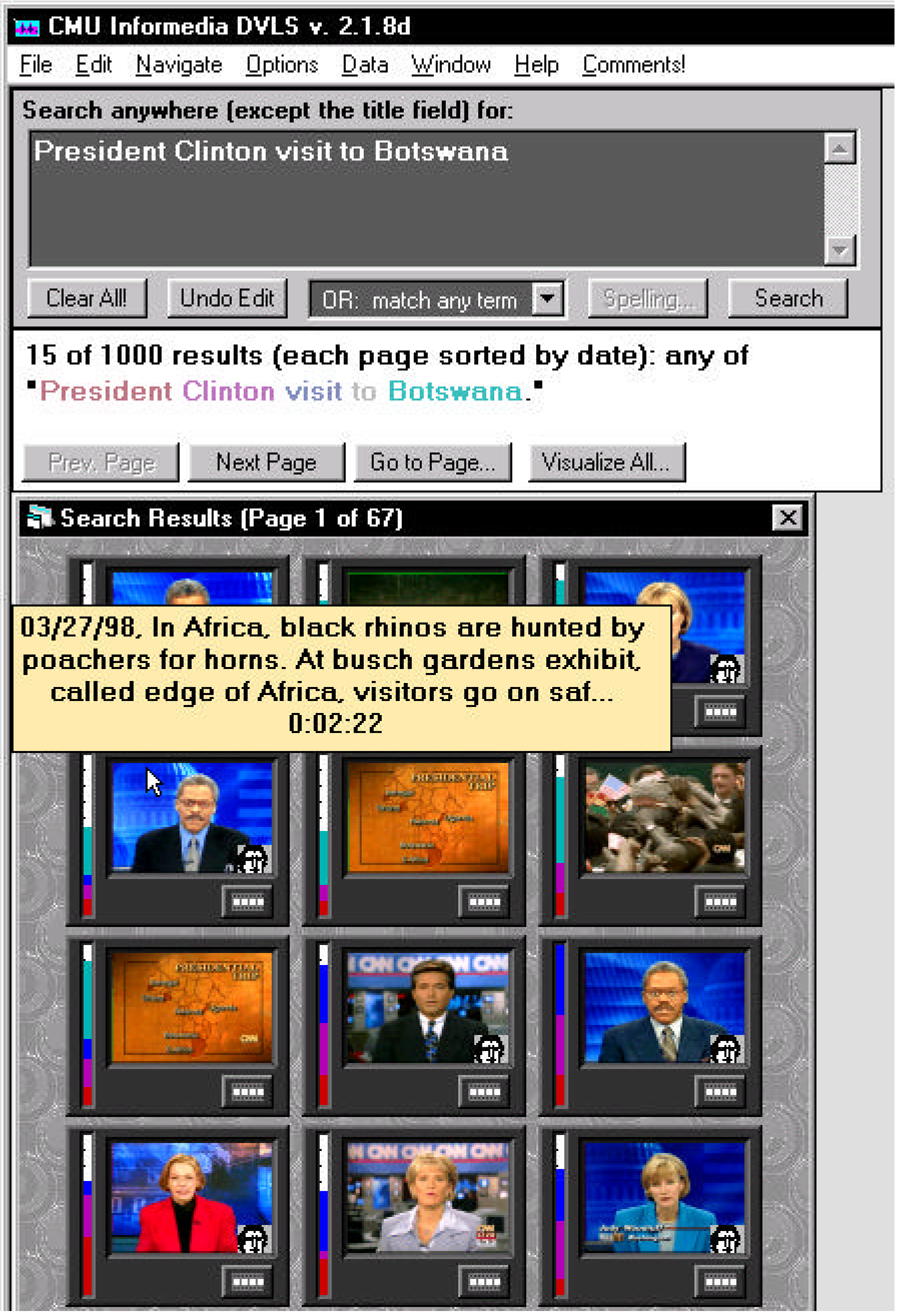
04/02/98, Syed abdul majid came to U.S. From ...

Figure 3. Seek to match location after a face search

2.2. Multimedia abstractions

A number of multimedia abstractions are implemented in IDVLS. The use of these abstractions has been discussed elsewhere (Christel et al., 1997b). However, an overview of these abstractions is appropriate here because they provide a more detailed visualization into a single video segment. These abstractions have been used primarily as aids in selecting segments from a list of search results for further inspection, and as a means to navigate within a selected segment.

Headlines are automatically generated for the IDVLS news corpus, using a heuristic to include more phrases from the start of a segment because news stories typically communicate more information up front. In the past IDVLS assembled the highly weighted words (using tf*Idf as a relevance measure) for a story, and limited the headline, or title, to 64 characters. Informal feedback indicated that the titles did not read well with this scheme, and so we changed titles to be an assemblage of the highly weighted phrases with a total length up to 128 characters. One such title (for the result in the second row, first column) is shown in Figure 4.



CMU Informedia DVLS v. 2.1.8d
File Edit Navigate Options Data Window Help Comments!
Clear All! Undo Edit OR: match any term Spelling... Search
Prev. Page Next Page Go to Page... Visualize All...

Search anywhere (except the title field) for:

President Clinton visit to Botswana

15 of 1000 results (each page sorted by date): any of "President Clinton visit to Botswana."

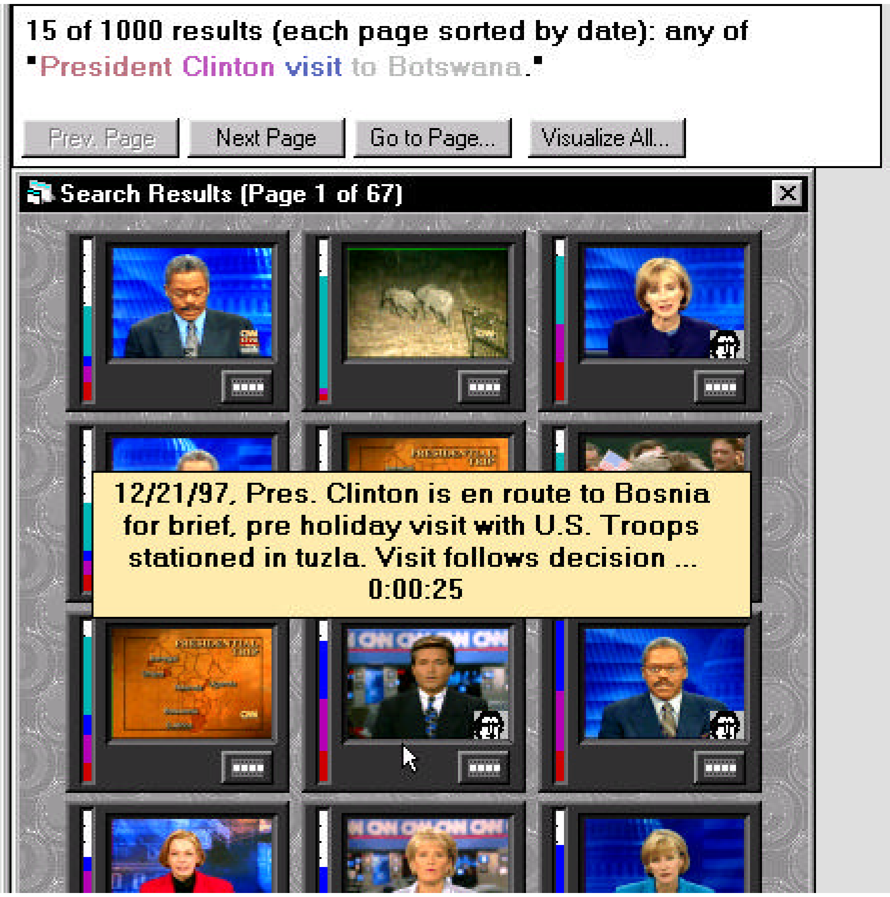
Search Results (Page 1 of 67)

03/27/98, In Africa, black rhinos are hunted by poachers for horns. At busch gardens exhibit, called edge of Africa, visitors go on saf...
0:02:22

Figure 4. Presentation of query results along with a title window for one of the results

The title shown in Figure 4 is preceded by the date "03/27/98" which was the date that story was shown. If the user wishes to have the results sorted by date, or results in each page sorted by date, then that date is given prominence by displaying it at the start of each title window, followed by the headline text and then the running time of the segment. If the user would be primarily interested in relevance then the title window would contain the headline text first, followed by the date and then the running time. This is one trivial case showing how multimedia abstractions can be tailored to usage context.

Each result in Figure 4 has a vertical relevance bar to the left of the thumbnail image. Taller non-white areas within this bar indicate more relevant queries. Through early 1998 IDVLS used a single relevance color, red, for all scoring words, and this bar was colored only with red. We are now experimenting with the use of color within this relevance bar as additional feedback to the user communicating why a particular result scored as it did. For example, the segment with its title shown matches on all four of "President", "Clinton", "visit" and "Botswana" but the primary scoring is due to the match on "Botswana", signified by the cyan color taking as much height as the other three colors combined.



15 of 1000 results (each page sorted by date): any of "President Clinton visit to Botswana."

12/21/97, Pres. Clinton is en route to Bosnia for brief, pre holiday visit with U.S. Troops stationed in tuzla. Visit follows decision ...
0:00:25

Figure 5. Title and query word coloring for another segment in result set (contrast with Figure 4)

Other feedback that has existed in IDVLS regardless of whether the scoring query words are all red or different colors is given in the results summary text appearing above the window with the thumbnail images. In Figure 4 you see four words colored with non-gray color in the phrase "President Clinton visit to Botswana." These four words are colored because they all were found in the result under the current mouse position, in this case the result in the second row first column. The word "to" appears in gray because it is such a common word that no matching is performed on it. If the mouse is moved over the result in the third row and middle column, then the word "Botswana" becomes gray because it is not found in that result, as shown in Figure 5. The word "visit" is shown in a more saturated blue because it contributes more to the relevance of that result than the one in Figure 4, where "visit" is shown in a less saturated blue. Hence, the coloring of the query words indicates both which words scored and the relative contributions of the words to the relevance score for that particular result. If red is used for all scoring words, the words with greater red saturation will have scored more than words with less saturation. Gray words always indicate words that are not found or words that are too common to be scored.

The single thumbnail image used to represent a segment in the result set is another form of multimedia abstraction. Prior Informedia papers have referred to this representative segment image as a "poster frame" (Christel et al., 1997a; Christel et al., 1997b). Early prototypes of IDVLS used poster frames as the default video abstraction to return to the user, as is done with the VISION library (Li et al., 1996). A formal study was conducted with 30 subjects during the summer of 1996 to determine whether a poster frame menu provides quicker access to relevant video clips than a text title menu. This study addressed the question of whether text titles are sufficient for presenting search results or if poster frames improved the interface. Results from the study showed that poster frames, when chosen based on the query, lead to significantly faster location of the relevant video by a user over the presentation of only a text title menu (Christel et al., 1997a). However, poster frames chosen arbitrarily, e.g., the first images in each video segment, do not offer any statistically significant advantages over a text title menu. Hence, if the poster frames shown in Figures 4 and 5 were chosen arbitrarily they may provide no performance improvements over simple text menus. However, by choosing the thumbnail images based on the usage context, i.e., based on the query producing the result set, the imagery provides significant performance and satisfaction improvements based on the results of the cited study (Christel et al., 1997a).

The poster frame can be tailored because each video segment typically is composed of several shots, with each shot represented by its own thumbnail image. The poster frame is taken to be the shot's thumbnail image where the highest scoring matches occur. Figure 6 shows the shot images for the segment selected in Figure 4; the display of all the shot break images at once has been referred to as the "filmstrip" video abstraction (Christel et al., 1997b). Colored notches above the shot images indicate where matches (either on words or on images) occurred. In this case, the matches are concentrated in the first shot showing the anchor. Hence, that shot is used to represent the segment in the result set as shown in Figure 4. Figures 4, 5, and 6 also show a face image overlaid in the bottom center or bottom right corner of the thumbnail images for those images where the automatic face detector recognized the existence of a face. The user can pick up and drag any of these images to the query window to initiate a face search using the dragged image as the search key.

Filmstrips date back to the advent of digital video (Mills et al., 1992). A number of researchers have converged on the idea of including an image in the filmstrip for each shot in the video (Aoki et al., 1996; Zhang et al., 1995a; Zhang et al., 1995b). Others suggest that more than one image should be added per shot depending on the composition of the shot determined through motion analysis (Wolf, 1996). Other researchers have implemented subsampling; in which a filmstrip image is extracted at evenly distributed intervals across a video (Taniguchi et al., 1995). The Informedia library makes use of subsampling as well as shot detection, object tracking, and other image processing techniques to determine which images to use in a filmstrip (Smith and Kanade, 1996). The second filmstrip image in Figure 6 is the result of subsampling: the image content did not change significantly but a duration threshold was reached and so another filmstrip image was added. The other images in Figure 6 are the result of new shots being detected because of image changes. Empirical testing needs to be conducted to verify that the selected frames represent the video source efficiently and serve the user effectively. CAETI is another digital video library making use of poster frames and filmstrips (Wolf et al., 1995).

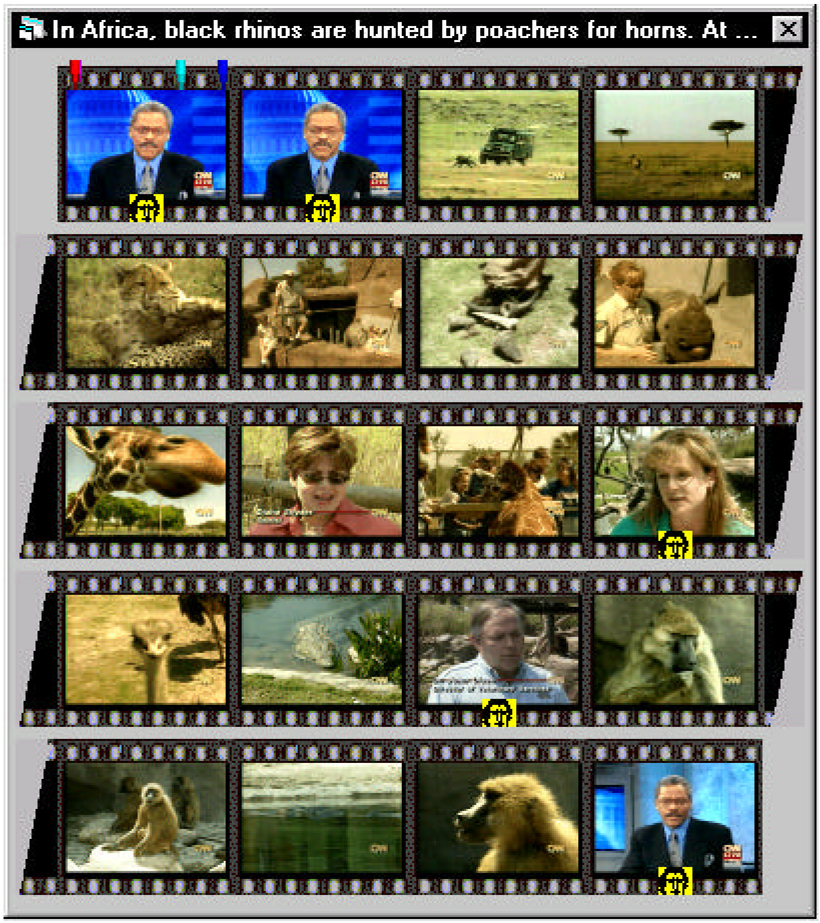

In Africa, black rhinos are hunted by poachers for horns. At ...

Figure 6. Filmstrip (shot thumbnail images) for the same segment selected in Figure 4

In Figure 6 the notches show where the matches occurred. In Figures 2 and 3 a line on the video ruler shows where the match occurred for those video segments. If multiple matches had occurred, then multiple lines would be drawn on the video ruler, color-coded to map back to query words if appropriate. Other techniques exist for abstracting the distribution behavior of query terms within a video clip, including TileBars (Hearst, 1995). Match lines on video and match notches on filmstrips were chosen for their simplicity and for their ability to facilitate navigation: by clicking on the match line or notch, the video's playback position is updated to the position corresponding to that match. The filmstrip itself provides another navigation mechanism as it communicates temporal structure in a static display. Considering Figure 6 again, if the user wishes to play back the giraffe sequence he or she can click on the giraffe image to jump directly to that point in the video segment.

"Video skims" are the most ambitious multimedia abstraction, in that they attempt to preserve the temporal nature of video, capture the essential content of a longer video and yet still achieve a 10 to 25% reduction in playback time (Christel et al., 1997b, Christel et al., 1998). Skims are played rather than viewed statically like filmstrips. Formal empirical investigations of skims were conducted using 72 subjects across two experiments. Results show that skims built from phrases rather than words, skims which preserve synchrony between audio and video, and skims with certain audio characteristics like breaking on silence pauses have the greatest potential for success (Christel et al., 1998).

These multimedia abstractions (headlines, poster frames, filmstrips, and skims) are in their own right a form of information visualization. They communicate a message embodied in the much larger source video. They can be manipulated to access that source video. They can be used to access other data, e.g., the text of the headline can be used to generate a word search or a shot break image containing a face can be used to initiate a face search. However, the information embodied in an abstraction is constrained to a single video document. The visualizations to be discussed in the next section communicate information about a set of video documents.

3. Information visualizers

Spoerri (1993) notes that queries, especially boolean queries, can be difficult for users to formulate and manage. Their results are often difficult to control, as anyone who has had to deal with web search engines knows when presented with thousands of web page hits in a linear list. The results are presented through a very limiting view (e.g., the top ten documents), and often, the user is given no helpful information as to how to modify the query to produce better result sets.

Visualizations of the digital video library data enable the library to no longer be merely queried but also to be navigated in unforeseen ways. The digital library interface expands from the "one-shot query" to address browsing, selecting, evaluating, thinking, and other information gathering and organizing tasks (Furnas, 1996). This section will briefly survey past information visualization work, and then discuss the techniques used to provide additional navigation and browsing capabilities to the Informedia interface.

3.1. Survey of past work

Robertson, Card and Mackinlay state that the basic goal for an "information appliance" is to "lower the cost of finding information and accessing it once found" (Robertson et al., 1993, p. 57). The same authors note that the traditional information retrieval paradigm needs to be expanded (p. 59): "in addition to concern with recall and precision, we also need to be concerned with reducing the time cost of information access and increasing the scale of information that a user can handle at one time." They propose four strategies to address these concerns:

1. Make user's immediate workspace larger.
2. Enable user interaction with multiple agents.
3. Increase the real-time interaction rate between the user and system.

4. Use visual abstraction to shift information to the perceptual system to speed information assimilation and retrieval.

In this paper we will concentrate exclusively on this fourth strategy.

Some visualization schemes have been proposed which build from the use of a query as the primary access into a corpus. For example, InfoCrystal uses Venn diagrams to visually represent the results of boolean queries and to act as a visual query language (Spoerri, 1993). Visualization By Example (VIBE) plots results as points against query words which move as the query words are moved by the user (Olsen et al., 1993). Judging from derivative systems, the gravitational metaphor of VIBE has proven more popular than InfoCrystal, with three-dimensional extensions to VIBE occurring in VR-VIBE (Benford et al., 1995) and Lyberworld (Hemmje et al., 1994).

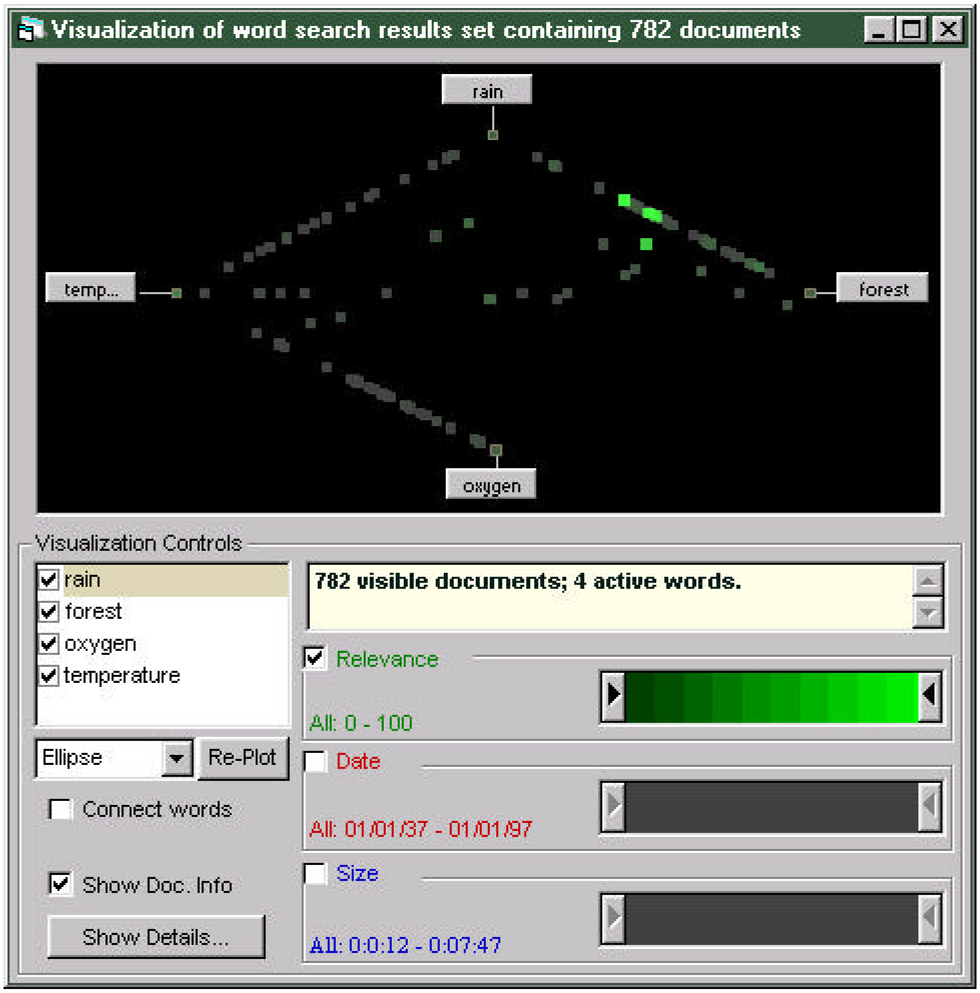
Other visualization schemes do not require a query as a starting point. The Spatial Paradigm for Information Retrieval and Exploration (SPIRE) creates graphic displays based on word similarities and themes in text (Wise et al., 1995; Risch et al., 1996). HomeFinder and FilmFinder use two-dimensional scatter plots for displaying the result sets or the complete corpus (Ahlberg and Shneiderman, 1994a and 1994b). Narcissus uses nodes and arcs with nodes possessing behavioral characteristics to attract and repel each other (Hendley et al., 1995). These techniques do not produce a single snapshot of a document base; rather, they provide views that can be manipulated to best meet a user's needs or perspective.

Some visualization schemes are particularly well suited to certain types of data. Within the Xerox PARC's Information Visualizer Cone Trees are used for hierarchical data, Perspective Walls for linear data, Data Sculptures for continuous data, and an office floor plan for spatial data (Robertson et al., 1993). Tree-Maps are a space-filling approach for displaying hierarchical information in the form of nested rectangles (Johnson and Shneiderman, 1991). Other techniques such as timelines (Kominek and Kazman, 1997), Media Streams (Davis, 1994), and Lifestreams (Freeman and Fertig, 1995; Freeman and Gelertner, 1996) focus on visualizations for the temporal structure of data. Fish-eye views (Furnas, 1986; Sarkar and Brown, 1994) can be used in conjunction with other techniques to present a focal area in greater detail while still preserving the context of the surrounding area.

The initial visualization technique added to IDVLS was based on VIBE. VIBE was chosen because it is well documented in the literature and has met with some success given that others have extended it into three-dimensional systems. The IDVLS user base is very familiar with querying the video library; VIBE extends the query metaphor well by enabling users to browse through whole result sets rather than linearly flip through result pages. We plan on adding more visualization capabilities in the future, e.g., if all video segments get classified into a topic hierarchy we could use Tree-Maps, Cone Trees, or a derivative to present that hierarchy. VIBE has proven to be a good candidate for introducing visualization into an interface that formerly had been dominated by query result examination.

3.2. A VIBE interface for video results

The initial VIBE implementation for IDVLS uses a two-dimensional scatter plot of rectangles. We chose to implement a two-dimensional presentation rather than three-dimensional because many of the gains offered by 3-D systems like Lyberworld deal with ambiguity reduction. However, by enabling manipulation of the scatter plot, the user can resolve ambiguities within the series of 2-D presentations. Figure 7 shows a VIBE plot following a query on “rain forest oxygen temperature”.



Visualization of word search results set containing 782 documents

782 visible documents; 4 active words.

☑ rain
☑ forest
☑ oxygen
☑ temperature

Ellipse

☐ Connect words

☑ Show Doc. Info

☑ Relevance
All: 0 - 100

☐ Date
All: 01/01/37 - 01/01/97

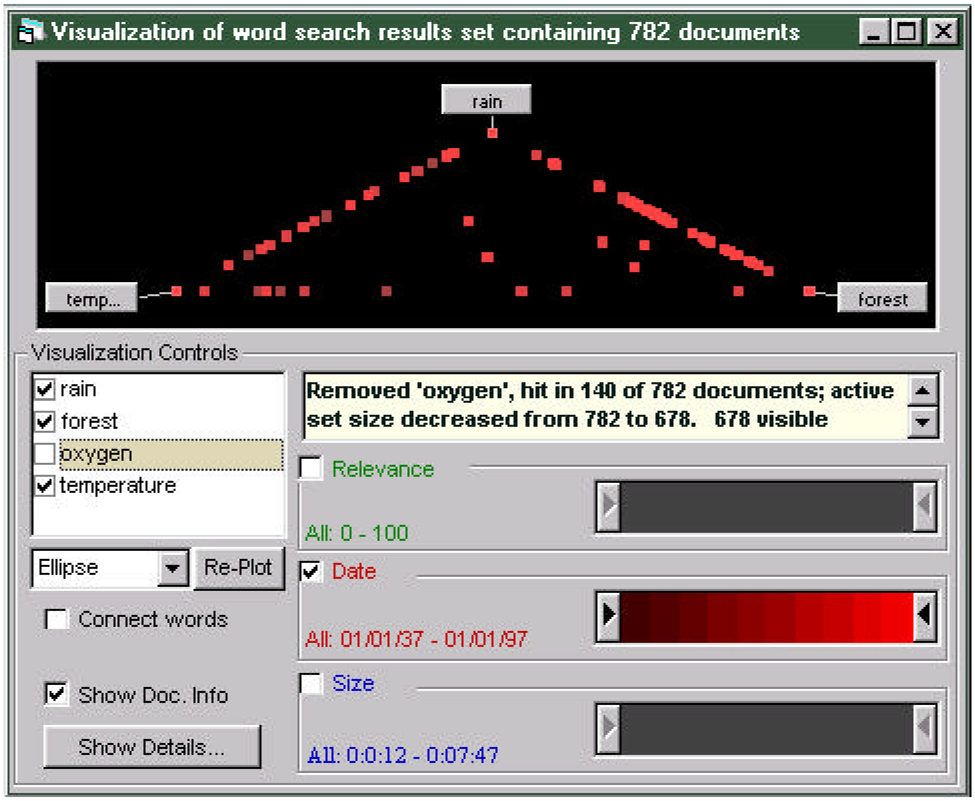
☐ Size
All: 0:0:12 - 0:07:47

Figure 7. IDVLS interface following "rain forest temperature oxygen" query

The video segments, generically referred to as “documents” in this interface, are displayed as colored rectangles, with color signifying relevance, date, and/or size. In Figure 7, only relevance is checked and active, so documents are colored as green rectangles, with a more saturated green indicating a higher relevance according to the scoring information provided by the word search engine. The location of a document is solely determined by the query words’ anchor positions and contributing scores for that document. For example, a document that has the same relevance score for “rain” and “forest” but which has zero score for the other words would be drawn exactly halfway between the anchor points for “rain” and “forest.” From this visualization, we can see that there are many relevant documents discussing “rain” and “forest” with a skew toward “forest”, less relevant documents discussing “oxygen” or “rain,” along with “temperature,” and very few documents discussing both “forest” and “oxygen” without at least one of “rain” or “temperature.”

If two documents are written to the same location, the green coloring for the rectangle placed there will be the combination of the documents' relevance scores. Thus, a bright green rectangle could indicate a high scoring document or numerous low scoring documents all at the same point. One method for resolving this ambiguity is zooming into the visualization; this may separate some of the overlapping documents. Another method lets the user mouse over the rectangles and be shown a message as to how many documents are represented by that rectangle (see Figure 13).

The interface supports direct manipulation in that the user can pick up and drag the query anchor words, as well as remove them from the display. In the example shown in Figure 7, unchecking "oxygen" removes that word from the display, as well as removing 104 documents that had matched only on "oxygen." In addition, 36 other documents that matched on "oxygen" as well as other checked query words are repositioned when oxygen is removed from the display. The result is shown in Figure 8, which also shows the documents as red boxes: a more saturated red color indicates a more recent document, with the dates ranging from 1937 to 1997.



☑ rain
☑ forest
□ oxygen
☑ temperature

□ Connect words
☑ Show Doc. Info

Removed 'oxygen', hit in 140 of 782 documents; active set size decreased from 782 to 678. 678 visible

□ Relevance
All: 0 - 100

☑ Date
All: 01/01/37 - 01/01/97

□ Size
All: 0:0:12 - 0:07:47

Figure 8. Results from Figure 7 after "oxygen" is removed from consideration and "Date" checked rather than "Relevance" for use in coloring the document representation

As an example of direct manipulation, suppose the user wanted to treat "rain" and "forest" as equals and focus on results that matched on temperature or oxygen and either "rain" or "forest." The user could drag "rain" and "forest" next to each other and position the other words at the opposite end of the visualization canvas, as shown in Figure 9.

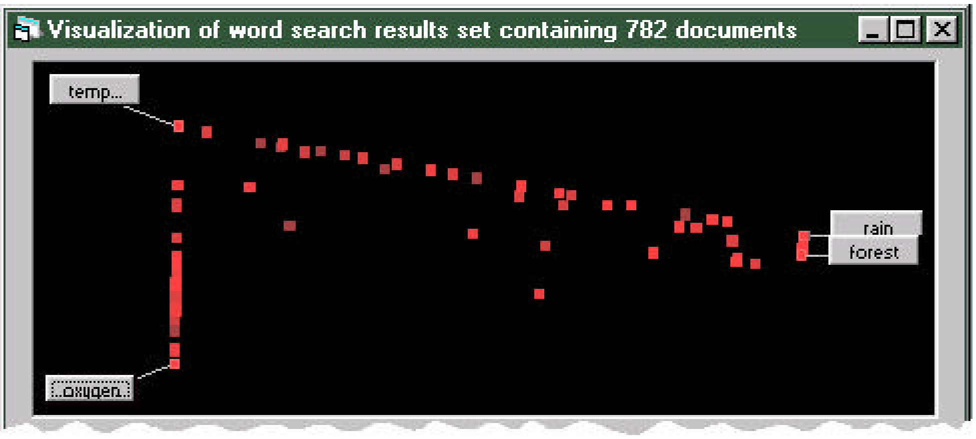
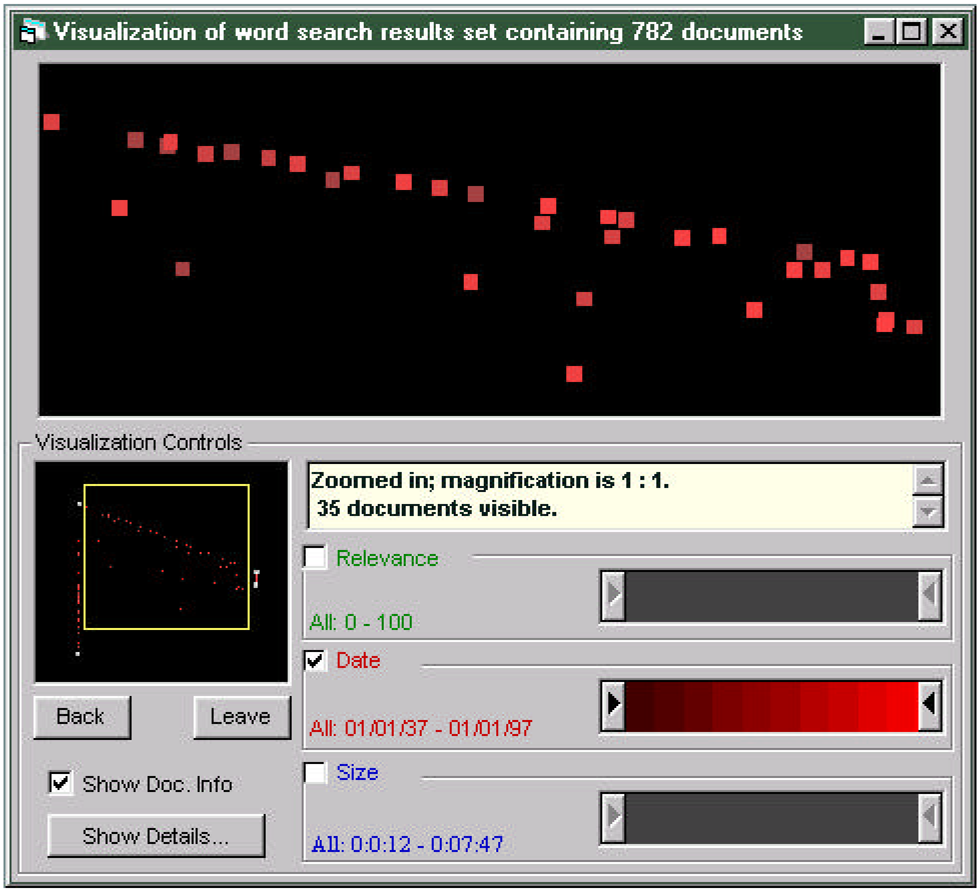

Visualization of word search results set containing 782 documents
temp...
rain
forest
oxygen

Figure 9. Example of repositioning query word anchors (see Figure 7 for normalized positions)


Visualization of word search results set containing 782 documents
Visualization Controls
Back
Leave
Show Details...

Zoomed in; magnification is 1 : 1.
35 documents visible.

☐ Relevance
All: 0 - 100

☑ Date
All: 01/01/37 - 01/01/97

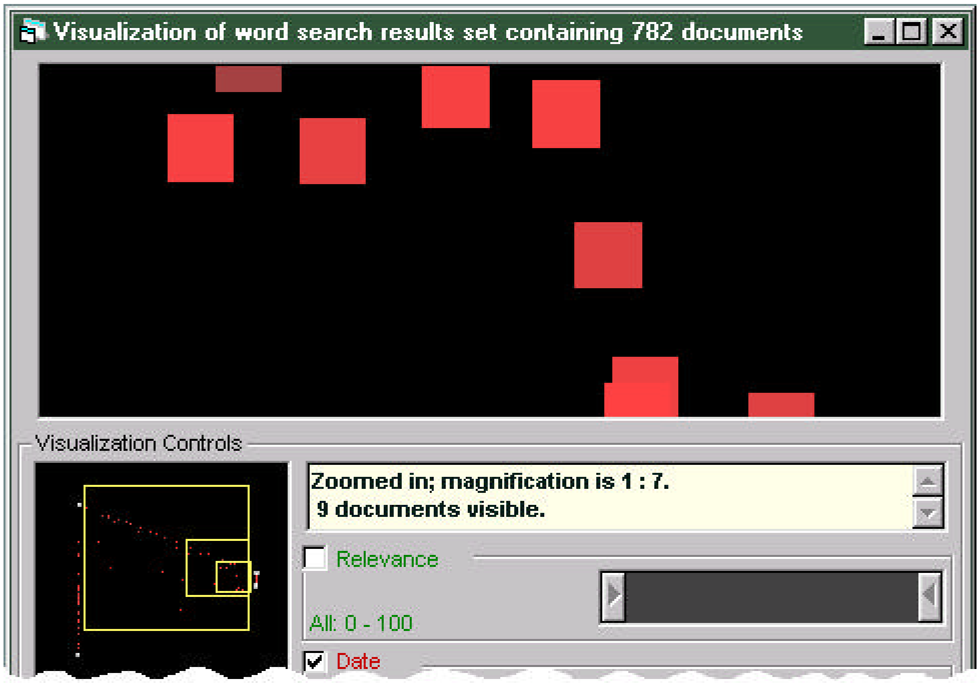
☐ Size
All: 0:0:12 - 0:07:47

☑ Show Doc. Info

Figure 10. Zoomed in view of subset of documents from Figure 9 layout

3.3. Using the interface for information filtering

Zooming is accomplished by dragging a selection rectangle in the visualization canvas. Figure 10 shows the results of zooming into the area between the query words shown in Figure 9. If no documents are contained within the selection, then no zoom operation is performed. The zoom context is preserved in the overview map shown in the lower left corner of the visualization window, as shown in Figure 10. The user can repeatedly zoom into the canvas, as shown in Figure 11. With each zoom, the rectangle representation for a document increases in size to give another form of feedback to the user.


Visualization of word search results set containing 782 documents
Visualization Controls

Zoomed in; magnification is 1 : 7.
9 documents visible.

□ Relevance

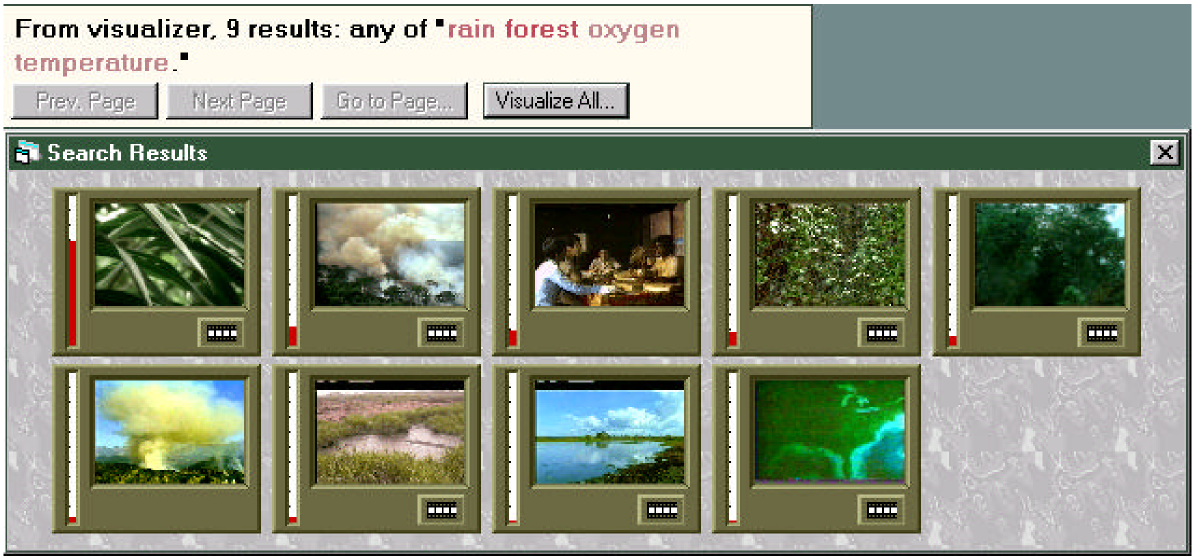
All: 0 - 100

☑ Date

Figure 11. Multiple zooms into the visualization canvas

To complete this example, the user may issue the query and not be satisfied to linearly look through 782 documents, displayed in thumbnail pages (see Figures 4 and 5) of up to 30 documents per page. Instead, they may use the visualizer as shown in Figures 7 through 11 to narrow down the results to those which match on either or both of rain/forest and either or both of oxygen/temperature. The number of qualifying documents has been reduced from 782 to 9, at which point the user may decide to “zoom into details” and show these 9 results with their title and thumbnail abstractions. By clicking on the “Show Details…” button in the visualizer interface, the display of Figure 12 is produced.

By combining the VIBE visualization technique with the presentation of thumbnail images and other details, IDVLS provides for a form of semantic zooming, where documents change their appearance as the amount of display real estate given to them changes (Furnas and Bederson, 1995). When completely zoomed out, the result space is fully aggregated and appears as Figure 7. By drilling down, detail is added to either the visualization as a whole (more attributes are shown; see discussion below), to the attributes in the view (e.g., text labels are added), or to a subset of the full result space. Figure 12 shows how details can be presented for a subset of the full result space. The user can then “zoom in” to a particular video segment or region within a segment by selecting to play that segment. This concept of "drilling down" or zooming in to add dimensions or detail to the display and zooming out or "rolling up" to aggregate and hide detail has been implemented successfully in systems like Visage (Roth et al., 1996) and Pad++ (Furnas and Bederson, 1995). The importance of preserving context while zooming into a document space has been empirically validated (Schaffer et al., 1996); such context preservation remains a design constraint for our work.



From visualizer, 9 results: any of "rain forest oxygen temperature."

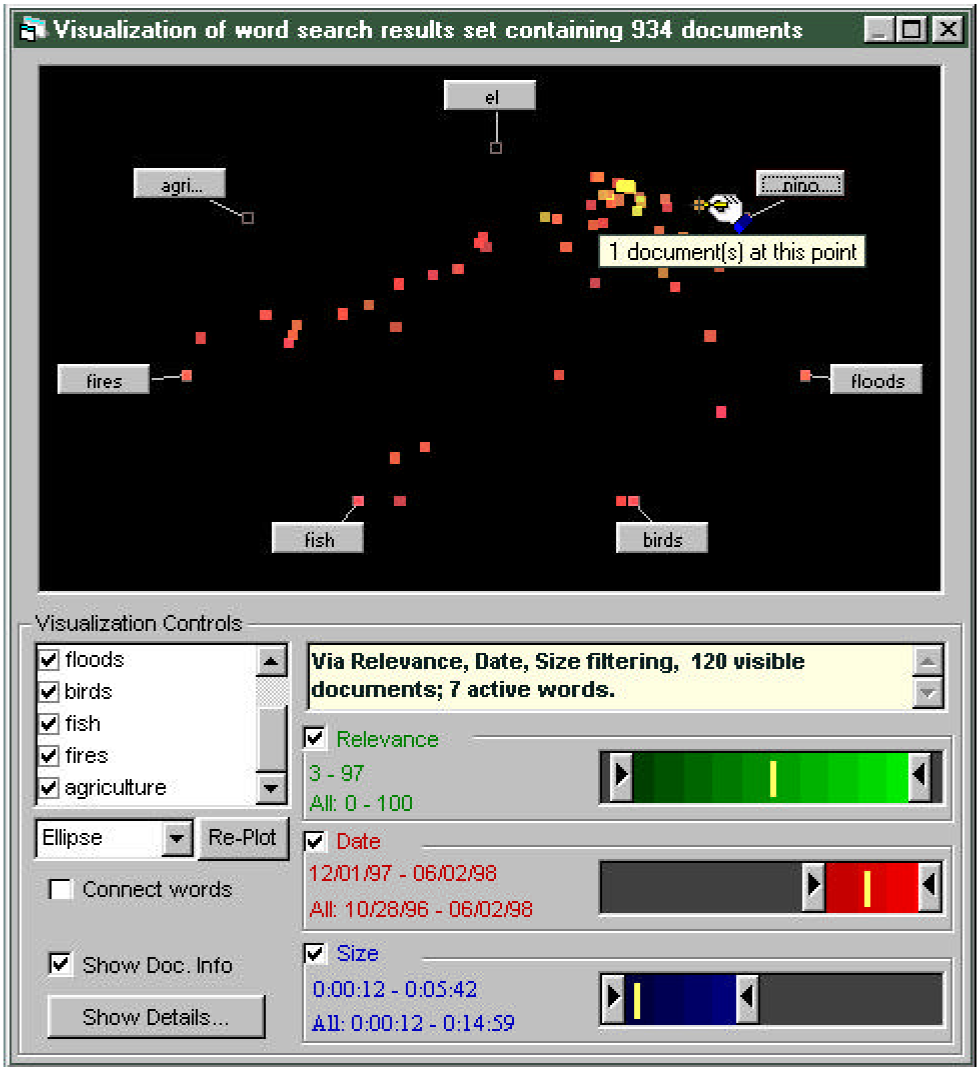
Figure 12. Details for subset of results defined by visualizer

3.4. Adding other information dimensions to the visualizer interface

The use of color can be exploited to convey further information about documents in the search result set to the user, with the RGB components of color providing three additional axes for data representation. This is illustrated in Figure 13 where the red, green and blue components of a document box's color represent its size, relevance (as returned by the search engine) and date respectively. Thus a box shown with a low intensity red component but high intensity green and blue components represents a small, recent document of high relevance to the query words.

The color sliders allow the user to rapidly and intuitively refine the search result set by selecting ranges for the red, green and blue intensity scales, discarding the documents that lie outside the selected scope. This use of sliders as dynamic query filters originates with Ahlberg and Shneiderman (1994a). Figure 13 presents dynamic slider settings that show only the most recent documents of small size, excluding the most and least relevant.

Initial prototyping showed that this overloading of color (when three color sliders were all being used simultaneously to represent the document) confused users. We addressed the problem in two ways: the user can disable, i.e., uncheck or turn off, the color sliders, and the user can see feedback on the slider concerning the document(s) under the mouse point. Figure 7 had the date and size sliders turned off and Figures 8 through 11 had the relevance and size sliders turned off. Figure 13 shows yellow marks on the three active sliders for the one yellow-orange box representing a document at the mouse point: these marks tell the user that the document at the mouse point is very small size, in the middle of the relevance scale, and fairly recent. The feedback could be extended to show distributions across size, time, and relevance for sets of documents or all the visible documents; such use of dynamic query sliders to communicate distributions has been done in the Influence Explorer (Tweedie et al., 1996). Another option is to explore the use of size and shape rather than just red, green, and blue saturation to represent other information dimensions.


Visualization of word search results set containing 934 documents

☑ floods
☑ birds
☑ fish
☑ fires
☑ agriculture

Ellipse

□ Connect words

☑ Show Doc. Info

Via Relevance, Date, Size filtering, 120 visible documents; 7 active words.

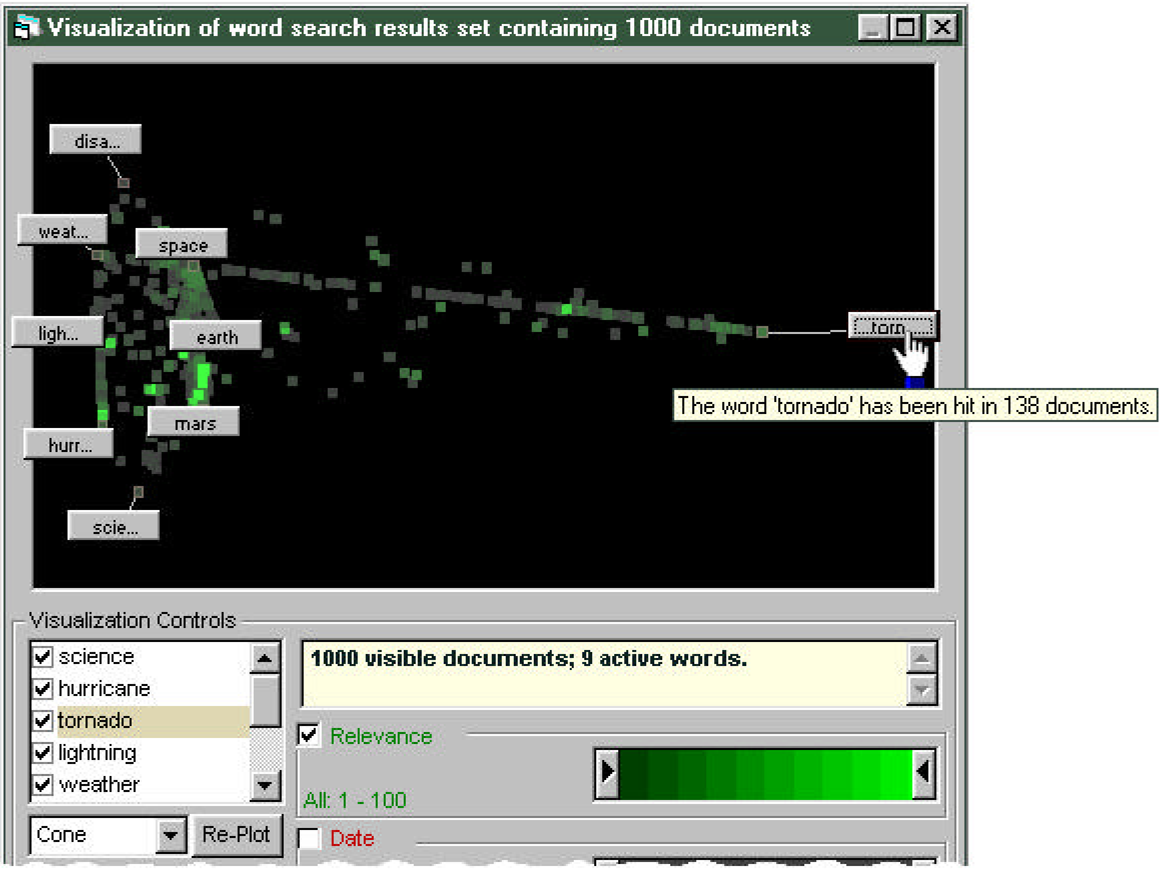
☑ Relevance
3 - 97
All: 0 - 100

☑ Date
12/01/97 - 06/02/98
All: 10/28/96 - 06/02/98

☑ Size
0:00:12 - 0:05:42
All: 0:00:12 - 0:14:59

Figure 13. IDVLS VIBE interface using green, red, and blue colors for relevance, size, and date, dynamic sliders, and feedback concerning the distribution of documents under the cross-hairs mouse

Also built into IDVLS VIBE are two preset positions of query words, which emphasize certain relationships between the documents in the search result set and the words. The first is the normalized “ellipse” position where each word is positioned at a vertex of a regular polygon, stretched to correspond to the display window dimensions (see Figures 7 and 13). This position gives the viewer an overall idea as to which documents are relevant to which query word. The second preset position is the “cone,” shown in Figure 14, in which the user selects the word to be used as the apex of the cone with the rest of the words forming the cone’s base. The cone structure is useful in separating out documents that are relevant to only the specified apex word. The concept of presetting word positions can be extended to show a variety of other query word relationships.


Visualization of word search results set containing 1000 documents

The word 'tornado' has been hit in 138 documents.

1000 visible documents; 9 active words.

☑ science
☑ hurricane
☑ tornado
☑ lightning
☑ weather

Cone

☑ Relevance

All: 1 - 100

□ Date

Figure 14. Preset IDVLS VIBE layout illustrating relationship of "tornado" to result set

4. Future directions

Users of the initial prototype of the work described in the previous section have been very positive with their comments. For example, one user searching on "weapons of mass destruction" wrote about the benefits of the "hits" feedback and direct manipulation of the IDVLS VIBE interface:

> I could find much more quickly with this feature which segment had the most use of these words in it, specifically the Nichols trial verdict and Sec. Cohen talking at length about the Iraqi situation. It would have been almost impossible to find with just the weights and thermometer visuals and paging through over 900 hits; neither of these was on the first page of results in the original query.

Future prototypes of the visualization interface will be refined through a series of evaluation techniques, including contextual inquiry, heuristic evaluation, cognitive walkthroughs, and think-aloud protocols. Graduate HCI students at Carnegie Mellon University (Lu et al., 1997) have applied these techniques to IDVLS successfully in the past. Other query and visualization interface references, especially those focused on digital libraries (Shneiderman et al., 1997), will be consulted to guide the further development of this work. We may conduct focused studies to verify the benefits of the visualization interface empirically, much as we did with query-based thumbnail images (Christel et al., 1997a) and different derivations of skims (Christel et al., 1998).

We are continuing to refine the current visualizer that begins with the result set from a query. We recognize the need for visualizers that work on the complete corpus in support of browsing without querying, the need to address dynamic corpora, the benefits of displays incorporating natural layouts for the data, and the opportunity to leverage from the unique visual and temporal nature of video. This concluding section touches upon some of these themes.

4.1. Interface improvements

Our visualization work has been motivated by the goal of semantic zooming. Currently there are three representation layers: the scatter plot high level view (as in Figure 13), the detailed subset view (as in Figure 12) and the single document view (as in Figure 2). We wish to pursue further “drilling down” to show more detail but perhaps less context (due to screen real estate) and "drilling up" to show more context but less detail, as done elsewhere in Visage (Roth et al., 1996) and Pad++ (Furnas and Bederson, 1995). Intermediate layers could be added, e.g., a layer between scatter plots and the detailed subsets in which size, more color, shape, and text labels could be used to show additional information dimensions, as is done in Envision (Nowell et al., 1996; Nowell et al., 1997). These attributes may not discriminate well at the scatter plot zoom level for hundreds of documents, but may work well for an order of magnitude fewer documents. At all zoom levels, fisheye views could be utilized to keep some context on the same screen as focused information while also addressing the limits of the display area.

We will continue to emphasize *interacting* with documents rather than just simply viewing documents. The current interface supports direct manipulation of query terms, relevance, date, and score scales and zooming areas. Other input displays need to be explored, e.g., the use of a labeled timeline to specify date ranges for information filtering. Other attributes may prove more useful for filtering. Based on trials with the current prototype, frequently repeated activities can be identified and the user interface streamlined to better accommodate those activities. For example, trials with the first IDVLS VIBE interface showed users frequently wish to inspect how one word affects the result space, which led to the cone word emphasis feature shown in Figure 14.

We recognize the importance of maintaining perceptual continuity. The points in the scatter plot move as the anchor query words are dragged. The context overview displays how multiple zooms are nested within each other (see Figure 11). The filmstrip button changes appearance as the mouse rolls over it in the detailed thumbnail image view (see Figure 4). Small cues like these reduce the cognitive costs of manipulating the interface. As more techniques are added the support for object constancy will remain a key issue. For example, if there are new ways of changing the user's perspective of the data, animation will be preferred over a jump cut.

4.2. Visualizing a dynamic corpus

The IDVLS VIBE interface begins with the results from a word query. We plan to investigate other techniques such as Themescapes (Wise et al., 1995) and Kohonen self-organizing maps (Chen et al., 1997) for navigating and browsing the corpus as a whole.

A significant problem exists when applying many of these techniques to a dynamic corpus, e.g., a news video corpus into which stories are added daily. A story that was located close to another in a current visualization may be repositioned to a completely different neighborhood when such additions occur. Some researchers suggest that the whole visualization not be reworked based on small additions, but that instead topological consistency should be preserved (Lueders and Ernst, 1995). In the Narcissus visualizer, each document has a behavior that repels it from some objects while attracting it to others (Hendley, 1995). With a dynamic corpus in Narcissus, the user could watch the visualization self-organize into a different layout, providing feedback concerning the changes brought on by yesterday's news, for example. Our first attempts at whole corpus visualizations will likely preserve views of the data at different times to enable tracking of a particular document over time as well as visualizing the changing nature of the corpus itself.

4.3. Visualizing to natural dimensions

Earlier figures showed scatter plots of documents mapped to the query words that produced a result set. In the absence of a query the information space needs to be arranged along other dimensions. Others have organized 2-D graphs of videos where the horizontal and vertical axes represented such attributes as date, producer, evaluation rating, copyright holder, and length (Nowell et al., 1997; Ahlberg and Shneiderman, 1994b). By allowing user selection of which features are mapped to the axes, and augmenting the graph with other tools like sliders, the user is able to generate and manipulate a multitude of views into the data.

Often, the data itself suggests an ideal spatial layout. In HomeFinder, a map of the Washington D.C. area is used to plot points representing homes in a real estate application (Ahlberg and Shneiderman, 1994a). Where appropriate, we plan on using maps for presenting digital video library data, such as with a library of news stories. The zoom levels could be set so that news is browsed on the world, continent, area, country, city, and neighborhood scopes. We will investigate automatically generating overviews, where individual video segments are intelligently aggregated into broader themes and perspectives. For example, one such overview could be highlighted areas on a world map showing locations of heavy flooding brought on by El Niño.

4.4. Incorporating video's temporal nature in visualizations

Each video in a digital library contains numerous streams of data such as people in the scenes, topics, locations, and dialogues. Timelines (Kominek, 1997) and Media Streams (Davis, 1994) both represent these individual streams as scrollable timelines or rectangular bars which are synchronized to the video, much like the text transcripts and filmstrips are synchronized to the video within IDVLS. We plan on extending our multimedia abstraction work to enable better visualization of the information contained within a video. Such work has increased importance given that new streams of information are being generated by the Informedia Project's suite of automatic processing tools, including captions from "video OCR" processing (Sato et al., 1998). Also, users can dynamically add their own annotations to video sequences within IDVLS.

By varying the segmentation in effect within the digital video library, another form of information zooming is provided. Time could be used as a means of aggregating information, so that points in a scatter plot could represent:

- video segments, as was done for this paper
- full source videos, e.g., the whole hour news show or two hour feature film
- video series, e.g., a week's worth of news or a twenty hour documentary series

A thousand hours of video consume over a terabyte of disk space. The key to accessing and interpreting this data need not and perhaps can not reside solely in an ideal search engine returning precisely the segment of interest. The user may not know enough about the corpus, the search engine, or his or her own needs to express an ideal query even if such an ideal engine existed. Appropriate visualization techniques give another window into the data, enabling users of the digital video library to "overview first, zoom and filter, then details-on-demand."

3. Acknowledgements

This material is based on work supported by the National Science Foundation, DARPA, and NASA under NSF Cooperative Agreement No. IRI-9411299, as well as work supported by DARPA through Contract No. N66002-97-C-8517. We thank WQED Pittsburgh, the U.K.'s Open University, the other Informedia Project partners for their support, and the Informedia Project team members for all of the underlying work in image processing, natural language processing, speech recognition, library data creation, library delivery, and project management. See the Informedia Project URL at http://www.informedia.cs.cmu.edu/ for complete lists of sponsors and team members.

4. References

Ahlberg, C. and Shneiderman, B. (1994a). Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays, *Proc. ACM CHI '94 Conference on Human Factors in Computing Systems*, Boston, 313-322.

Ahlberg, C. and Shneiderman, B. (1994b). The Alphaslider: A Compact and Rapid Selector, *Proc. ACM CHI '94 Conference on Human Factors in Computing Systems*, Boston, 365-371.

Aoki, H., Shimotsuji, S., and Hori, O. (1996). A Shot Classification Method of Selecting Effective Key-Frames for Video Browsing, *Proc. ACM Multimedia Conference*, (Boston, MA, November), 1-10.

Benford, S., Snowdon, D., Greenhalgh, C., Ingram, R., Knox, I., and Brown, C. (1995). VR-VIBE: A Virtual Environment for Co-operative Information Retrieval, *Proc. Eurographics '95*, (Maastricht, The Netherlands, August/Sept.), 349-360.

Chen, H., Schatz, B., Houston, A., Sewell, R., Ng, T., and Lin, C. (1997). Internet Browsing and Searching: User Evaluations of Category Map and Concept Space Techniques, *Proc. 2nd ACM International Conference on Digital Libraries*, (Philadelphia, PA, July), 257.

Christel, M., Stevens, S., Kanade, T., Mauldin, M., Reddy, R., and Wactlar, H. (1996). Techniques for the Creation and Exploration of Digital Video Libraries. Chapter 8 of *Multimedia Tools and Applications*, B. Furht, ed. Kluwer Academic Publishers.

Christel, M. G., Winkler, D. B., & Taylor, C. R. (1997a). Improving Access to a Digital Video Library, *Human-Computer Interaction INTERACT '97: IFIP TC13 International Conference on Human-Computer Interaction*, (Sydney, Australia, July), 524-531.

Christel, M. G., Winkler, D. B., & Taylor, C. R. (1997b). Multimedia Abstractions for a Digital Video Library. *Proc. 2nd ACM International Conference on Digital Libraries*, (Philadelphia, PA, July), 21-29.

Christel, M. G., Smith, M. A., Taylor, C. R, and Winkler, D. B. (1998). Evolving Video Skims into Useful Multimedia Abstractions, *Proc. of the ACM CHI '98 Conference on Human Factors in Computing Systems*, (Los Angeles, CA, April).

Davis, M. (1994). Knowledge Representation for Video, *Proceedings of AAAI '94*, 120-127.

Freeman, E., and Fertig, S. (1995). Lifestreams: Organizing your Electronic Life, *AAAI Fall Symposium: AI Applications in Knowledge Navigation and Retrieval*, November, Cambridge, MA.

Freeman, E., and Gelernter, D. (1996). Lifestreams: A Storage Model for Personal Data, *ACM SIGMOD Bulletin*, March.

Furnas, G. (1986). Generalized fisheye views, *Proceedings of ACM CHI '86 Conference on Human Factors in Computing Systems*, (April), 16-23.

Furnas, G. and Bederson, B. (1995). Space-Scale Diagrams: Understanding Multiscale Interfaces, *Proceedings of the ACM CHI'95 Conference on Human Factors in Computing Systems*, (Denver, CO, May), 234-241.

Furnas, G. (1996). User Interface Design Panel at Digital Library Initiative All-Project Meeting. Ann Arbor, MI, USA, May 16-17, 1996. See http://www.si.umich.edu/UMDL/aui.html for more recent perspectives from the UMDL Advanced User Interface team.

Gong, Y. (1998). *Intelligent Image Databases: Towards Advanced Image Retrieval*. Hingham, MA: Kluwer Academic Publishers.

Hauptmann, A. G., Witbrock, M. J., and Christel, M. G. (1995a). News-on-Demand - An Application of Informedia Technology, *D-LIB Magazine*, September, URL http://www.dlib.org/dlib/september95

Hauptmann, A. G., Witbrock, M. J., Rudnicky, A. I., and Reed, S. (1995b). Speech for Multimedia Information Retrieval, *UIST-95 Proceedings*, (Pittsburgh, November), 79-80.

Hauptmann, A. G. (1998). Video indexing and speech recognition in digital libraries. *Proc. IEEE Advances in Digital Libraries*, (Santa Barbara, April).

Hearst, M. A. (1995). TileBars: Visualization of Term Distribution Information in Full Text Information Access, *Proceedings of the ACM CHI'95 Conference on Human Factors in Computing Systems*, (Denver, CO, May), 59-66.

Hemmje, M, Kunkel, C., and Willett, A. (1994). LyberWorld -- A Visualization User Interface Supporting Fulltext Retrieval, *Proc. 17th Int'l Conf. on Research and Development in Information Retrieval (SIGIR '94)*, 249-259.

Hendley, R. J., Drew, N. S., Wood, A.M., and Beale, R. (1995). Narcissus: Visualising Information, *Proceedings of IEEE Symposium on Information Visualisation*, Atlanta.

Johnson, B., and Shneiderman, B. (1991). Tree-Maps: A Space-Filling Approach to the Visualization of Hierarchical Information Structures. *Proc. IEEE Visualization '91*, (San Diego, October), 284-291.

Kominek, J., and Kazman, R. (1997). Accessing Multimedia through Concept Clustering, *Proceedings of ACM CHI '97 Conference on Human Factors in Computing Systems*, (Atlanta, March), 19 - 26.

Li, W., Gauch, S., Gauch, J., and Pua, K. M. (1996). VISION: A Digital Video Library, *Proceedings of the ACM Conference on Digital Libraries*, (Bethesda, MD, March), 19 - 27.

Lu, F., Rosen, D., and Spitznagel, B. (1997). Unpublished HCI evaluation of the Informedia Digital Video Library System, fall semester.

Lueders, P. and Ernst, R. (1995). Research Report: The Automatic Display Layout: Improving Information Visualization, *Proc. IEEE Information Visualization '95*, Los Alamitos, CA.

Marchionini, G. (1995). *Information Seeking in Electronic Environments*. Cambridge Series on Human Computer Interaction. Cambridge University Press.

Mills, M., Cohen, J., and Wong, Y. Y. (1992). A Magnifier Tool for Video Data, *Proceedings of the ACM CHI'92 Conference on Human Factors in Computing Systems*, 93-98.

Nowell, L. T.; France, R. K.; Hix, D.; Heath, L. S.; Fox, E. A. (1996). Visualizing Search Results: Some Alternatives to Query-Document Similarity, *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, (Zurich, Switzerland, August), 67-75.

Nowell, L., France, R., and Hix, D. (1997). Exploring Search Results with Envision, *Extended Abstracts of ACM CHI '97 Conference on Human Factors in Computing Systems*, (Atlanta, March), 14 - 15.

Olsen, K. A., Korfhage, R. R., Sochats, K. M., Spring, M. B., and Williams, J. G. (1993). Visualization of a Document Collection: The VIBE System. *Information Processing & Management*, **29**(1), 69-81.

Risch, J. S.; May, R. A.; Dowson, S. T.; Thomas, J. J. (1996). A Virtual Environment for Multimedia Intelligence Data Analysis, *IEEE Computer Graphics and Applications*, November, 33-40.

Robertson, G., Card, S., and Mackinlay, J. (1993). Information Visualization Using 3D Interactive Animation. *Communications of the ACM* **36**(4), 56-71.

Roth, S., Lucas, P., Senn, J., Gomberg, C., Burks, M., Stroffolino, P., Kolojejchick, J., & Dunmire, C. (1996). Visage: A User Interface Environment for Exploring Information, *Proc. IEEE Info. Visualization '96*, 3-12.

Rowley, H., Baluja, S., and Kanade, T. (1995). Human Face Detection in Visual Scenes, Carnegie Mellon University, *CS Technical Report CMU-CS-95-158*, Pittsburgh, PA.

Sarkar, Manojit and Brown, Marc H. (1994). Graphical Fisheye Views. *Comm. of the ACM* **37**(12), 73 - 84.

Salton, G. (1989). *Automatic Text Processing.* Reading, MA: Addison-Wesley Publishing Company, Inc.

Sato, T., Kanade, T., Hughes, E., and Smith, M. (1998). Video OCR for Digital News Archives, *Proc. IEEE Workshop on Content-based Access of Image and Video Databases*.

Schaffer, D., Zuo, Z., Greenberg, S., Bartram, L., Dill, J., Dubs, S., and Roseman, M. (1996). Navigating Hierarchically Clustered Networks Through Fisheye and Full-Zoom Methods. *ACM TOCHI Transactions on Computer-Human Interaction* **3**(2), 162-188.

Shneiderman, B. (1996). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. Human Computer Interaction Laboratory, Institute for Systems Research, Institute for Advanced Computer Studies, *Dept. of Computer Science Tech. Report CS-TR-3665*, Univ. of Maryland, July.

Shneiderman, B., Byrd, D., and Croft, W. B. (1997). Clarifying Search: A User-Interface Framework for Text Searches. *D-LIB Magazine*, URL http://www.dlib.org/dlib/january97

Smith, M. and Kanade, T. (1996). Video Skimming for Quick Browsing Based on Audio and Image Characterization, Carnegie Mellon University, *CS Technical Report CMU-CS-95-186R*, Pittsburgh, PA.

Spoerri, A. (1993). InfoCrystal: A Visual Tool for Information Retrieval and Management, *Proceedings of the 2nd International Conf. on Information and Knowledge Management (CIKM93),* Washington, D.C., 11-20.

Taniguchi, Y., Akutsu, A., Tonomura, Y., and Hamada, H. (1995). An Intuitive and Efficient Access Interface to Real-Time Incoming Video Based on Automatic Indexing, *Proceedings of the ACM Multimedia Conference*, (San Francisco, CA, November), 25-33.

Tweedie, L., Spence, R., Dawkes, H., and Su, H. (1996). Externalising Abstract Mathematical Models, *Proceedings of ACM CHI '96 Conference on Human Factors in Computing Systems*, (April), 406-412.

Wactlar, H. D., Kanade, T., Smith, M. A., and Stevens, S. M. (1996). Intelligent Access to Digital Video: Informedia Project. *Computer*, **29**(5), 46-52.

Wise, J. A., Thomas, J. J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., and Crow, V. (1995). Visualizing the non-visual: Spatial analysis and interaction with information from text documents, *Proc. IEEE Information Visualization '95,* Los Alamitos, CA, 51-58.

Wolf, W., Liu, B., Wolfe, A., Martonosi, M., and Liang, Y. (1995). A Digital Video Library for Classroom Use, *Proceedings of the International Symposium on Digital Libraries 1995*, (Tsukuba Science City, Japan, August), 250-255.

Wolf, W. (1996). Key Frame Selection by Motion Analysis, *Proceedings of the 1996 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* (Atlanta, May), 1228-1231.

Zhang, H. J., Smoliar, S. W., Wu, J. H., Low, C. Y., and Kankanhalli, A. (1995a). A Video Database System for Digital Libraries, in *Digital Libraries: Current Issues (Digital Libraries Workshop DL '94, Newark, NJ, May 1994, Selected Papers)* (eds. N.R. Adam, B.K. Bhargava, and Y. Yesha), Springer, Berlin.

Zhang, H.J., Low, C.Y., and Smoliar, S.W. (1995b). Video parsing and browsing using compressed data. *Multimedia Tools and Applications*, **1**, 89-111.