

1-2006

Evaluation and User Studies with Respect to Video Summarization and Browsing

Michael G. Christel

Carnegie Mellon University, christel@cs.cmu.edu

Follow this and additional works at: <http://repository.cmu.edu/compsci>

Published In

Multimedia Content Analysis, Management and Retrieval 2006, Proceedings of IS&T/SPIE Symposium on Electronic Imaging, 6073,
.

This Conference Proceeding is brought to you for free and open access by the School of Computer Science at Research Showcase @ CMU. It has been accepted for inclusion in Computer Science Department by an authorized administrator of Research Showcase @ CMU. For more information, please contact research-showcase@andrew.cmu.edu.

Evaluation and User Studies with Respect to Video Summarization and Browsing

Michael G. Christel, School of Computer Science, Carnegie Mellon University,
Pittsburgh, PA, USA 15213; christel@cs.cmu.edu, phone 1 724 935-4076

ABSTRACT

The Informedia group at Carnegie Mellon University has since 1994 been developing and evaluating surrogates, summary interfaces, and visualizations for accessing digital video collections containing thousands of documents, millions of shots, and terabytes of data. This paper surveys the Informedia user studies that have taken place through the years, reporting on how these studies can provide a user pull complementing the technology push as automated video processing advances. The merits of discount usability techniques for iterative improvement and evaluation are presented, as well as the structure of formal empirical investigations with end users that have ecological validity while addressing the human computer interaction metrics of efficiency, effectiveness, and satisfaction. The difficulties in evaluating video summarization and browsing interfaces are discussed. Lessons learned from Informedia user studies are reported with respect to video summarization and browsing, ranging from the simplest portrayal of a single thumbnail to represent video stories, to collections of thumbnails in storyboards, to playable video skims, to video collages with multiple synchronized information perspectives.

Keywords: user studies, information visualization, digital video library, video surrogate, video collage, interface evaluation, video skim

1. INTRODUCTION

As the creation of digital imagery and video proliferates and as automated content-based video processing techniques improve, a wealth of visual materials are now available to end users. Concept-based strategies where annotators carefully describe digital photographs and video with text concepts that can later be used for searching and browsing are powerful but expensive. Estimates from the Library of Congress place the cost of professionally annotating an image at \$100, with a consulting expert confirming the cost of annotating another professional image collection at \$75-\$100 per image. Users have shown that they are unlikely to invest the time and labor to annotate their own photograph and video collections with text descriptors. Prior evaluations have shown that annotators do not often agree on the concepts used to describe the materials, so the text descriptors are often incomplete.

To address these shortcomings in concept-based strategies, content-based strategies work directly with the syntactic attributes of the source materials in an attempt to derive indices useful for subsequent browsing and retrieval. For video, the most common syntactic features are color, texture, shape, and coarse audio attributes such as speech/music or male/female speech. These lowest level content-based indexing techniques can be automated to a high degree of accuracy, but unfortunately in practice they do not meet the needs of the user, reported often in the multimedia information retrieval literature as the semantic gap between the capabilities of automated systems and the users' information needs. Pioneer systems like IBM's QBIC demonstrated the capabilities of color, texture, and shape search, while also showing that users wanted more.

Continuing research in the video information indexing and retrieval community attempts to address the semantic gap by automatically deriving higher order features, e.g., indoor/outdoor, face, people, cityscape, and greenery. Rather than leave the user only with color, texture, and shape, these strategies give the user control over these higher order features for searching through vast corpora of materials. The NIST TRECVID video retrieval evaluation forum has provided a common benchmark for evaluating such work, charting the contributions offered by automated content-based processing as it advances.

To date, TRECVID has confirmed that the best performing interactive systems for news and documentary video leverage heavily from the narration offered in the audio track. The narration is transcribed either in advance for closed-captioning by broadcasters or as a processing step through automatic speech recognition (ASR). In this manner, text concepts for concept-based retrieval are provided for video, without the additional labor of annotation from a human viewer watching the video, with the caveat that the narration does not always describe the visual material present in the video. Despite the lower accuracy of these text descriptions from the narrative, they still are the best source of indexing information, offering greater utility than the automated content-based techniques dealing with visual and non-speech aural features.

Because the text from narration is not as accurate as a human annotator describing the visual materials, and because the latter is too expensive to routinely produce, subsequent user search against the video corpus will be imprecise, returning extra irrelevant information, and incomplete, missing some relevant materials. Summarizing the returned results with video surrogates can help the user to be able to quickly and accurately weed out the irrelevant information and focus attention on the relevant material, addressing precision. The term “document surrogate” is used in the information retrieval community to label information that serves as the representation for the full document, such as a title, abstract, table of contents, set of keywords, or combinations of these descriptors. In this paper “video surrogate” is used to label the set of text, image, audio, and video that can serve as a condensed representation for the full video document.

As for the problem in recall, that the returned results miss some relevant materials, providing the user with information visualization interfaces allows broad chunks of the corpus to be browsed efficiently. Information visualization enables the user to conduct additional investigations in addition to specific queries, investigations that allow exploration into information regions of interest selected by the user. Such exploration has the potential to turn up additional relevant items that would not be found through the specific queries.

Video retrieval may be in this state for some time, with human text concept-based tagging of video materials too laborious, incomplete, and expensive, and with automated content-based indexing resulting in a semantic gap when at an accurate but syntactic level, and being imperfect with many sources of error when attempted at higher order semantic levels. The user interface is critical to enabling the user to wade through increasing amounts of video information and locate materials of interest in light of imprecise and incomplete indexing strategies.

Since 1994, the Informedia research group at Carnegie Mellon University has been developing and testing numerous interfaces for accessing terabytes of video, including work on surrogates that represent a video document or set of video documents in an abbreviated manner. The Informedia collections contain primarily CNN broadcast news dating back ten years, but also other U.S., Chinese, and Arabic news broadcasts, documentaries, interviews, and surveillance footage. Overall, over ten terabytes of video has been processed, with the news video alone consisting of nearly 200,000 story segments and over 3 million shots. This paper walks through some of the Informedia user studies conducted through the years, discussing the evaluation and evolution of surrogates such as thumbnails, storyboards, video skims, and video collages. The merits of discount usability techniques for iterative improvement and evaluation are discussed, as well as the structure of formal empirical investigations that address the human computer interaction metrics of efficiency (can I finish the task in reasonable time), effectiveness (can I produce a quality solution), and satisfaction (would I be willing or eager to repeat the experience again). The three metrics may be correlated, e.g., an interface that is very satisfying may motivate its user to greater performance and hence higher effectiveness, while conversely an unsatisfying interface may produce extremely slow activity leading to poor efficiency. These three usability aspects are discussed elsewhere in greater detail as they relate to HCI research in general, with the conclusion that all three are necessary to get an accurate assessment of an interface’s usability¹. Before surveying the Informedia user studies, a discussion of ecological validity is warranted, because it affects the impact of the user study results. Foraker Design defines ecological validity as follows²:

Ecological validity – the extent to which the context of a user study matches the context of actual use of a system, such that it is reasonable to suppose that the results of the study are representative of actual usage and that the differences in context are unlikely to impact the conclusions drawn. All factors of how the study is constructed must be considered: how representative are the tasks, the users, the context, and the computer systems?

Ecological validity is often difficult for multimedia information retrieval researchers for a number of reasons. The data in hand may not be representative, e.g., the use of the Corel professional image database will not be represent amateur collections like the average individual's digital photograph collection. The tasks employed may be artificial, e.g., finding a factual date from a news video corpus may be a task that in practice is always achieved through a newspaper text archive rather than a broadcast news archive. The users may not represent actual users, with university research often substituting college students as the user study subjects because of their availability. Finally, the context is likely different between the user study and an actual work environment, with an actual work environment having time and accuracy pressures that are difficult to simulate in a short term study. A discussion of ecological validity will be threaded throughout the survey of Informedia user studies.

2. THE BENEFITS OF THUMBNAILS AND QUERY CONTEXT

Video is an expensive medium to transfer and view. MPEG-1 video, the compressed video format used in most of the Informedia collections, consumes 1.2 Megabits per second. Looking through an hour of candidate video for relevant material could take an hour of viewing time and require downloading over 500 Megabytes of information. Surrogates can help users focus on precisely which video documents are worth further investigation and where to focus attention within those documents, reducing viewing and video data transfer time.

Consider a user interested in the western African countries from a corpus of January-June 2005 news. The query produces 75 results in the Informedia library of CNN news video during this period. Figure 1 shows a scrolling window of the results, where each document is represented by a brief title and a single thumbnail image overview. As the user moves the mouse cursor over a document representation, its title is displayed in a pop-up menu. The layout of Figure 1 communicates the relative relevance of each document to the query as determined by the map search engine, the contribution of each query term (in this case, countries) for each document (i.e., which countries matched which documents and by how much), a contextual thumbnail image representation, a brief title automatically produced for the document, and the document's play length and broadcast date.



Figure 1. Results from map search, with each video story result represented by a single thumbnail.


The vertical bar to the left of each thumbnail indicates relevance to the query, with color-coding used to distinguish contributions of each of the query terms. The document surrogate under the mouse cursor, the fifth result, has its title text displayed in a pop-up window, and the query word display is also adjusted to reflect the document under the

cursor³. From the query on the six shown countries, this fifth document matches only Somalia. The vertical relevance bar (to the left of each thumbnail) shows that this document has a relevance score of 25/100 for the given query. Glances at the relevance bar shows that “Egypt” in purple dominates the top results. The third result, with most of the score thermometer in purple for “Egypt” but a bit in red for “Libya”, is the only one in the top 12 matching on two of the query countries. Other interfaces with a temporal element, such as the storyboard interface and video playback window discussed further below, add views reflecting the distribution of these match terms within the video.

The utility and efficiency of the layout shown in Figure 1 have been reported in detail elsewhere^{3,4,5}, validated through a number of usability methods, including transaction log analysis, contextual inquiry, heuristic evaluation, and cognitive walkthroughs. Formal studies allow facets of surrogate interfaces to be compared for statistically significant differences in dependent measures such as success rate and time on task. In particular, a formal empirical study was conducted to determine the relative merits of such thumbnail menus of results versus similar text menus of titles, document durations and broadcast dates⁶. Thirty high school and college students participated in an experiment using a fact-finding task against a documentary video corpus, where dependent measures included correctness, time to complete the task, and subjective satisfaction. The study had high ecological validity because such students typically are shown documentaries and later asked to recall presented facts or produce reports based on the information within the documentaries. As such, questions on who would benefit and why from this study are easily answered: high school and college students would benefit, because finding material from documentary videos could be made more effective, efficient, and satisfying.

The study found that when the thumbnail image is chosen based on the query context, users complete the task more quickly and with greater satisfaction with such an interface than when using plain text menus containing no imagery, or when using a context-independent thumbnail menu, in which each document is always represented by the same thumbnail selection strategy of taking the first shot in the document. A simple thumbnail selection strategy did not distinguish itself from a straight text presentation. However, if the thumbnail to represent a video document is chosen based on where the query terms appear with the greatest score (a combination of density of matches and importance of matches as returned by the text search, image search, map search, or whatever search engine was used), then that query-based thumbnail does produce faster, more accurate, more satisfying retrieval performance. As an example of a query-based thumbnail, the Somalia story as the fifth result in Figure 1 starts off with an interview head shot in a studio that is much less informative visually than the brownish street shot of Somalia shown in Figure 1, with the street shot chosen based on the user’s query and query engine attributing more of the matches (to the specified western Africa geographic region) to the street shot than the interview shot. This result, that surrogates chosen based on context produce a more efficient visual interface, will be confirmed again and again in follow-up Informedia user studies.

3. THE BENEFITS OF STORYBOARDS AS VISUAL OVERVIEWS

The automatic breakdown of video into component shots has received a great deal of attention by the image processing community^{7, 8, 9, 10, 11}. TRECVID has had a shot detection task charting the progress of automatic shot detection since 2001, and has shown it to be one of the most realizable tasks for video processing with accuracy in excess of 90%¹². In video retrieval, a broadcast is commonly decomposed into numerous shots, with each shot represented by a keyframe: a single bitmap image, i.e., thumbnail, extracted from that shot. The numerous keyframes can then be subjected to image retrieval strategies. The thumbnail images for each shot can be arranged into a single chronological display, a storyboard surrogate, which captures the visual flow of a video document along with the locations of matches to a query. From Figure 1’s interface, clicking on the filmstrip icon () for a document displays a storyboard surrogate like that of Figure 2, with triangle notches at the top of thumbnails communicating some match context: what matched and where for a given query against this selected video.

The storyboard interface is equivalent to drilling into a document to expose more of its visual details before deciding whether it should be viewed. Storyboards are also navigation aids, allowing the user to click on an image to seek to and play the video document from that point forward. For example, the mouse is over the purple match notch of the second shot shown in Figure 3, a notch corresponding to the location EGYPT corresponding to this story’s inclusion in the map query results shown in Figure 1. If the mouse is clicked here, the corresponding video is opened and played from that point 1:30 into the document where Egypt is mentioned and when the aerial view shown in the second shot thumbnail is visible. Storyboard displays of a simultaneous, ordered set of thumbnail images date back to the advent of digital video.

Numerous commercial and research systems such as CAETI, EUROMEDIA, Físchlár, VideoLogger, and our own Informedia have implementations of storyboards showing keyframes arranged in chronological order^{13,14}.

Figure 2. Storyboard surrogate for the third video document of Figure 1.

Storyboards were found to be an ideal roadmap into a video possessing a number of shots, e.g., the user can quickly navigate to the video of the building in the fourth shot of Figure 2 by clicking on that fourth shot, rather than linearly playing the video to that point. Of course, for some video like an hour video of a single person talking, the whole video is a single shot of that person's head, and a storyboard of that one shot provides no navigational value. When there is a multiplicity of shots, storyboards can be very effective.

Hence, a major difficulty with storyboards is that there are often too many shots to display in a single screen^{6, 8, 20}. In Video Manga^{20, 21}, the interface presents thumbnails of varying resolutions, with more screen space given to the shots of greater importance. In the Informedia storyboard interface, the thumbnails are kept the same size with the lesson of the thumbnail-query context study applied to this situation: the user's query context can indicate which shots to emphasize in an abbreviated display. Rather than show all the shots, only those shots containing matches (i.e., those marked with match notches as shown for 5 of the 27 shots in Figure 2) can be included in a representation for a collection of video, so that rather than needing to show 1849 shots, 345 matching shots could be shown to represent the 75 segments given the query context of west Africa locations.

4. TEMPORAL SURROGATES: STORYBOARDS WITH TEXT, AND VIDEO SKIMS

While storyboard surrogates represent the temporal dimension of video, they do so in a static, visual-only manner. Transitions and pace may not be captured, and audio is not directly represented. The idea behind an Informedia "video skim" is to capture the essence of a video document in a collapsed snippet of video, e.g., representing a ten minute video as a one minute video skim that serves as an informative summary for that longer video. Skims are highly dependent on genre: a skim of a sporting event might include only scoring or crowd-cheering snippets, while a skim of a nursing home surveillance video might include only snippets where people are moving in the scene. Skims of educational documentaries were studied in detail by Informedia researchers, where users accessed skims as a comprehension aid to understand quickly what a video was about. They did not use skims for navigation, e.g., to jump to the first point in a nutrition documentary where salt is discussed. Storyboards serve as much better navigation aids because there is no temporal investment that needs to be made by the user, whereas for skims, the user must play and watch the skim.

For documentaries, the audio narrative contains a great deal of useful information. Early attempts at skims did not preserve this information well. Snippets of audio for an important word or two were extracted and stitched together in a skim, which was received poorly by users based on discount usability techniques, much like early text titles in the Informedia interface (e.g., one title is shown as tooltips text in Figure 1) comprised of the highest TF-IDF words were rejected in favor of more readable concatenated phrases. By extracting audio snippets marked by silence boundaries, the audio portion of the skim became greatly improved, as the skim audio was more comprehensible and less choppy.

A formal study was conducted to investigate the importance of aligning the audio with visuals from the same area of the video, and the utility of different sorts of skims as informative summaries²². The experimental procedure had each subject experience each treatment in a Latin Square design to counterbalance the ordering/learning effects, i.e., it was a within-subjects design. Five treatments were seen by each of 25 college students, as illustrated in Figure 3:

- DFS: a default skim using short 2.5 second components, e.g., comprising seconds 0-2.5 from the full source video, then seconds 18.75-21.25, seconds 37.5-40, etc.
- DFL: a default skim using long 5 second components, e.g., seconds 0-5, then seconds 37.5-42.5, 75-80, etc.
- NEW: a new skim, outlined here but discussed in more detail in the study paper²²
- RND: same audio as NEW but with reordered video to test synchronization effects
- FULL: complete source video, with no information deleted or modified

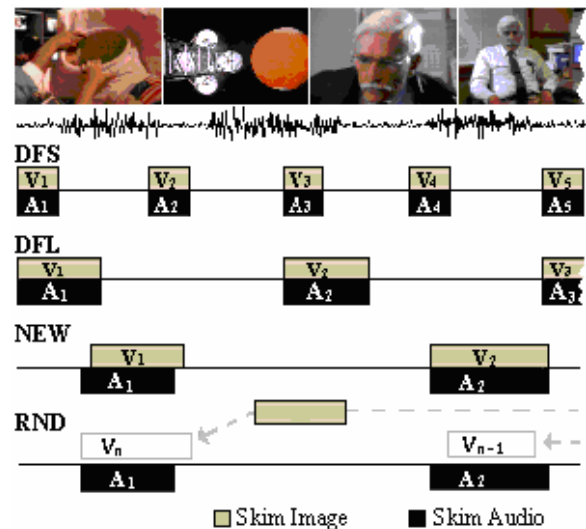


Figure 3. Skim treatments used in empirical study on skim utility as informative summary.

Specifically, it was believed that skims composed of larger snippets of dialogue would work better than shorter snippets, the equivalent of choosing phrases over words. A new skim was developed that comprised of snippets of audio bounded by significant silences, more specifically audio signal power segmentation. The transcript text for the audio snippets was ranked by TF-IDF values and the highest valued audio snippets were included in the skim, with the visual portion for the skim snippets being in the close neighborhood of the audio. Following a playing of either a skim or the full video, the subject was asked which of a series of images were seen in the video just played, and which of a series of text summaries would make sense as representing the full source video. As expected, the FULL treatment performed best, i.e., watching the full video is an ideal way to determine the information content of that full video. The subjects preferred the full video to any of the skim types. However, subjects favored the NEW skim over the other skim treatments, as indicated by subjective ratings collected as part of the experiment. These results are encouraging, showing that incorporating speech, language, and image processing into skim video creation produces skims that are more satisfactory to users. The larger component size, when used with signal-power audio segmentation, produced the NEW skim that did distinguish itself from the other skims. If the larger component size is used only for subsampling, however, it (DFL) yields no clear objective or subjective advantage over short component size skims (DFS). In fact, both DFS and DFL often rated similarly to RND, indicating perhaps that any trivial subsampled skim, regardless of granularity, may not do notably well. While very early Informedia skim studies found no significant differences between a subsampled skim and a “best” audio and video skim, this study uncovered numerous statistically significant differences²². The primary reasons for the change can be traced to the following characteristics of the audio data in the skim:

- Skim audio is less choppy due to setting phrase boundaries with audio signal-processing rather than noun-phrase detection.
- Synchronization with visuals from the video is better preserved.
- Skim component average size has increased from three seconds to five.

Usage data, HCI techniques, and formal experiments led to the refinement of single document video surrogates in the Informedia digital video library over the years. Thumbnail images are useful surrogates for video, especially as indicative summaries chosen based on query-based context. The image selection for thumbnails and storyboards can be improved via camera motion data and corpus-specific rules. For example, in the news genre shots of the anchorperson in the studio and weather reporter in front of a map typically contribute little to the visual understanding of the news story. Such shots can be de-emphasized or eliminated completely from consideration as single image surrogates or for inclusion in storyboards.

Again depending on genre, text can be an important component of video surrogates. Indeed, Ding et al. found that surrogates including both text and imagery are more effective than either modality alone¹³, confirmed in an Informedia user study which specifically examined the questions of text layouts and lengths in storyboards²³. 25 university students and staff members participated in an experiment using a fact-finding task against a news video corpus, where dependent measures included correctness, time to complete the task, and subjective satisfaction. In news video, information is conveyed both through visuals (especially field footage) and audio (such as the script read by the newscaster), so a mixed presentation of both synchronized shot images and transcript text extracts was expected to offer benefits over image-only storyboards. Significant differences in performance time and satisfaction were found by the study. If interleaving is done in conjunction with text reduction, to better preserve and represent the time association between lines of text, imagery and their affiliated video sequence, then a storyboard of images plus text with great utility for information assessment and navigation can be constructed. That is, the transcript text should be time-aligned with thumbnail rows in the storyboard, and then reduced to a set of basic phrases important to the particular query context. As with the video skim study, the conclusion from the storyboard-plus-text study shows that assembling surrogates from phrases (longer chunks) works better than assembling from words (shorter chunks), with synchronization between text, audio, and/or visuals very important. Integrating cues from multiple modalities can improve multimedia summarization interfaces.

Showing distribution and density of match terms is useful, and can naturally be added to a storyboard (Figure 2’s notches) or a video player’s play progress bar. The interface representation for the match term can be used to navigate quickly to that point in the video where the match occurs. Returning again to the point on ecological validity, however,

and we see that the real value for video surrogates and summarization is in addressing sets of video documents rather than navigating and summarizing a single one. Users lost in the space of a single hour document may sacrifice an hour to locate material of interest, but users lost in the space of a thousand hour video set cannot possibly find what they are after with reasonable performance, speed, or satisfaction: the utility of surrogates for summarizing sets of video increases dramatically.

5. INFORMATION VISUALIZATION: SUMMARIZING ACROSS SETS OF VIDEO DOCUMENTS

Traditionally, a query to a digital library produces a linear list of result documents. Locating the meaningful information in the results list becomes problematic because: (1) too much information is returned; (2) the list view neither communicates the meaning of the list as a whole nor the multiple relationships between items in the list; and (3) different users have different information needs. Informedia researchers developed the *video collage* as an interface for users to more quickly interpret and assimilate information relevant to their needs. A video collage is defined as an automatic presentation of text, images, audio, and video derived from multiple video sources in order to summarize, provide context, and communicate aspects of the content for the originating set of sources. Instead of sequencing through lists of query results, users can explore the video library through multiple video collages such as timelines emphasizing time, maps emphasizing geographic distribution, and storyboards of faces emphasizing associated people. Video collages can adapt dynamically based on user and usage information. Users can drill down into collages and see smaller subsets or see the contributions of individual documents to the summary presentation. Users can expand the collage to show more context, displaying large portions of the whole video library. Users can also discover trends and produce follow-up multimodal queries directly through interaction with the collages.

Figure 4a shows a timeline video collage for the results from a query on “James Jeffords” against a 2001 news library. The vertical axis is query relevance; the horizontal axis is broadcast date. The most common phrases, people, organizations and places for the 28 video documents returned by the query are automatically populated in four text list boxes beneath the timeline plot. The source text is derived from transcripts generated by speech recognition, closed captioning sources, overlay text extracted through image processing, and other automatic processing which may contain errors. The automated named entity extraction to identify people, places and organizations in the text metadata is also imperfect. Hence, the metadata incorporated into the collage contains errors, but despite those errors, the collage interface of Figure 1 has been shown to have summarization value for describing newsworthy people when compared to web biographical sketches²⁴. The study detailed below examines to what degree collages built from automatically derived data are effective when used by people to address their information needs. Through interaction with dynamic query sliders²⁵, the text descriptors, images, layout and scale in the collage change to reflect a more focused view on a smaller set of video documents. By adjusting the date slider to the crowded time period holding many Jeffords stories in late May, the Figure 4a presentation changes to that shown in Figure 4b.

When video collages were first developed and discussed, an evaluation compared them to other information summarization sources²⁴, showing that the collages’ text contents were reasonable summaries. This sort of an evaluation was straightforward, in that the text from the different sources could be directly compared using standard information retrieval metrics of precision and recall. What was missing, however, was an empirical study dealing with end users. In order to assess the value of text and value of thumbnail imagery as components in collages, a within-subjects experiment was conducted with 20 university students using four versions of timeline collages: with text (the 4 lists of Figure 4), with imagery (the thumbnails in Figure 4), with both text and imagery (as shown in Figure 4), and with neither text nor imagery (green dots rather than thumbnails are plotted in the timeline, with no text lists). The task was to complete a celebrity report, where 24 celebrities were chosen from the infoplease.com site for “2001 People in the News” as was done in a prior text-centric study without users that examined the text information shown by collages against other web-based sources²⁴.

The library for the study was 232 hours of CNN daily and weekend broadcasts from 2001, at least 30 minutes per day, segmented through closed-captioning into 11,595 video documents. Through speech, language, and image analysis, additional metadata was automatically generated, such as transcript timing, shots for each document with a representative image for each shot, identification of anchorperson and weather shots, and recognition of text overlaid on the video^{3, 24}. The experiment looked at the use of collages built from such metadata, specifically, 161,885 thumbnail

images and over 2 million words (mostly transcript words but also other categories like overlaid text) for this CNN 2001 study corpus.



Figure 4. Timeline collage from “James Jeffords” query against 2001 news (a), zooming into May 23-31 (b).

The task was chosen to represent the broad fact-gathering work supported by information visualization interfaces. Prior work with high school and college students and a digital video library showed that assignments frequently centered on the tasks of assembling answers to “who,” “what,” “when,” “where,” and “why” questions, along with creating visually appealing cover pages communicating the main theme of a report. Pilot tests were conducted with college students and staff against this particular news corpus and the 24 celebrities to trim down the set of items on a celebrity report template to those that could be answered successfully, without much ambiguity or redundancy, but also without reducing the task to a trivial exercise. The resulting report template kept text slots for “who,” “what,” “when,” and “where” responses, as well as image slots for portrait and cover shots representing the newsworthiness of that celebrity for 2001.

The usability of the collage interfaces was measured by including the recommended metrics for efficiency, satisfaction, and effectiveness¹. Efficiency was taken as the time to complete the celebrity report. Satisfaction was measured with a closing questionnaire asking subjects to remark on certain interface aspects, rank their treatment preferences, and provide whatever free-form text comments they wished to share. Effectiveness was measured through automatic and manual means. The precision and recall of the subject’s text answers were automatically graded based on the infoplease.com “2001 People in the News” web page for that celebrity. Precision addresses whether the words in the report are correct, i.e., the correct words divided by the total number of words in the report. Recall addresses whether the words in the report are complete, i.e., the number of correct words in the report divided by the total number of correct words in the truth, in this case taken to be the InfoPlease page. Because the InfoPlease web page for a celebrity may not represent truth well, the extraction of words from the answers is a coarse filter that may lose the original text’s meaning, and the matching of words is unforgiving for different word forms and synonyms, the automatic measures were supplemented with human assessment of the celebrity reports. Three human assessors graded the reports without knowledge of which treatments were used.

Subjects were able to successfully complete celebrity reports that earned high marks from the human graders for cover and portrait images, and who, what, when, and where text content across all four interface treatments. The text precision suffered for treatments having thumbnail images. Collages without images produced a significantly higher

precision in text answers than did collages with images, $F(1, 191) = 4.49, p < 0.05$. There was no significant difference for recall. The efficiency suffered as well, with more time taken to complete reports using collages with images, $F(1, 191) = 5.2, p < 0.03$. Despite problems with the thumbnails, subjects clearly favored them, $F(1, 76) = 5.9, p < 0.02$. When thumbnails were not present, some subjects clicked the scatterplot dots and played the videos represented in the collage in chronological order, left-to-right, irrespective of relevance. These same users, when the collage contained thumbnails, chose thumbnails that were visually striking or spatially isolated on the timeline, rather than in strict left-to-right order as with the dots. Again, the presence of imagery in the collage was found to directly affect the interaction patterns. No conclusions were found with respect to the presence of additional text lists in the collages. They were rarely accessed for copying or dragging into the report, despite often holding the answers to report fields. Users commented that the use of the text lists was not apparent.

As a direct result of this experiment, video collages were improved in the following ways: the text was integrated better with the rest of the presentation through “brushing”: interactions in the plot highlight and change the text lists and vice versa. Also, the thumbnails were plotted with a better layout strategy so as to overlap less frequently and communicate more visual information. The improvements were verified through follow-up discount usability techniques, but the significance and impact are debatable based on ecological validity: how natural is it to consult a broadcast news video corpus to learn details about a celebrity? How important is assimilation across stories, as provided by video collages, versus just reporting the information from the top-rated story from a text query against the corpus using the celebrity’s name? Who actually uses news video corpora, and for what purpose? Dealing with these questions as a single research institution still leaves possible numerous critiques. Is the automated processing tuned to solve the experimental task but not real-world tasks? Is the input video data a small but unrepresentative set and results will differ when dealing with real-world sized corpora? Is the input video data itself also tuned for success with certain tasks, e.g., eliminating news advertisements by hand ahead of time because they are noisy and mess up presentations (which was not done, by the way – just an example)? These questions argue for a community-wide forum for evaluating video retrieval interfaces and determining ecological validity, which brings us to a discussion of TRECVID.

6. TRECVID AS AN EVALUATION FORUM FOR VIDEO SUMMARIZATION AND BROWSING

The Text REtrieval Conference (TREC) was started in 1992 to support the text retrieval industry by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. At that time, the Cranfield tradition of using retrieval experiments on test collections was already well-established, but progress in the field was hampered by the lack of easily accessible, realistically large test collections. Large test collections did exist, but they were proprietary, with each collection usually the result of a single company’s efforts. The proprietary nature of the collections also biased them in various ways. TREC was conceived as a way to address this need for large, unbiased test collections. The same needs for the video retrieval community led to the establishment of the TREC Video Track in 2001. Now an independent evaluation, TRECVID began with the goal to promote progress in content-based retrieval from digital video via open, metrics-based evaluation. The corpora have ranged from documentaries to advertising films to broadcast news, with international participation growing annually¹². A number of tasks are defined in TRECVID, including shot detection, story segmentation, semantic feature extraction, and information retrieval.

The Cranfield paradigm of retrieval evaluation is based on a test collection consisting of three components: a set of documents, a set of information need statements called topics, and a set of relevance judgments. The relevance judgments are a list of the “correct answers” to the searches: the documents that should be retrieved for each topic. Success is measured based on quantities of relevant documents retrieved, in particular the metrics of recall and precision. The two are combined into a single measure of performance, average precision, which measures precision after each relevant document is retrieved for a given topic. Average precision is then itself averaged over all of the topics to produce a mean average precision (MAP) metric for evaluating a system’s performance.

For TRECVID video searches, the individual “documents” retrieved are shots, where a shot is defined as a single continuous camera operation without an editor’s cut, fade or dissolve – typically 2-10 seconds long for broadcast news. The TRECVID search task is defined as follows: given a multimedia statement of information need (a topic) and the common shot reference, return a ranked list of up to 1000 shots from the reference which best satisfy the need. Three types of search are studied: automatic in which the query topic is taken as is with no human modifications; manual in

which a human can rephrase the query topic into a form suitable for the specific system but after issuing the query interacts no further; and interactive in which the user can view the topic, interact with the system, see results, and refine queries and browsing strategies interactively while pursuing a solution. The interactive user has no prior knowledge of the search test collection or topics.

To address ecological validity, the topics are defined by NIST to reflect many of the sorts of queries real users pose, based on query logs against video corpora like the BBC Archives and other empirical data^{12, 26}. The topics include requests for specific items or people and general instances of locations and events, reflecting the Panofsky-Shatford mode/facet matrix of specific, generic, and abstract subjects of pictures²⁷.

User studies conducted with TRECVID topics on TRECVID data have a vast head start over studies like the collage study detailed earlier because they can make use of the TRECVID community effort to claim ecological validity in most regards: the data set is real and representative, the tasks (topics) are representative based on prior analysis of BBC and other empirical data, and the processing efforts are well communicated with a set of rules for all to follow. The remaining question of validity is whether the subject pool represents a broader set of users, with university students and staff for the most part comprising the subject pool for many research groups because of their availability. Over the years, Informedia TRECVID experiments have confirmed the utility of storyboards showing matching thumbnails across multiple video documents¹⁹, the differences in expert and novice search behavior when given TRECVID topics²⁸, the utility of transcript text for news video topics²⁹, and the overlooking of using feature filters (e.g., include or exclude all shots having the face feature or “outdoors” feature) to reduce the shot space^{19, 28, 29}.

The latter result led to a follow-up study investigating why concept-based filters have failed thus far to produce better TRECVID topic performance. A survey was given to 12 university employees and students asking them to map 10 TRECVID 2004 features to the 23 TRECVID 2004 topics, and also to map the 17 features to the 24 topics in TRECVID 2003. The survey set up the problem as follows: suppose there are tens of thousands of video shots and you need to answer a particular topic. You don’t have time to look through all the shots, but can choose to either look at or ignore shots having a feature. For example, if the topic were “cherry trees” you might decide to definitely look at outdoor shots and vegetation, and definitely ignore Madeleine Albright shots. The judgments showed the difficulty in assessing the utility of a feature for a given topic, with people often disagreeing on the relevance of a feature to a particular topic, including disagreement within the 8% of positive feature-topic associations strongly supported by truth data³⁰.

Other video browsing and retrieval lessons learned from Informedia TRECVID experiments include the utility of packing storyboards with visually dense presentations collapsed to match neighborhoods, and incorporating domain-specific content-based retrieval strategies (e.g., for news, eliminating anchors, emphasizing shots in middle of broadcast segments, separating news from commercials, identifying reporters, etc.). TRECVID provides a public corpus with shared metadata to international researchers, allowing for metrics-based evaluations and repeatable experiments. Its advantages for video retrieval user studies are further detailed elsewhere²⁹. An evaluation risk with over-relying on TRECVID is tailoring interface work to deal solely with the genre of video in the TRECVID corpus, for example dealing specifically with just U.S. news, the TRECVID corpus in 2004. This risk is mitigated by varying the TRECVID corpus genre: in 2005 it held Chinese and Arabic news as well as U.S. news, and in 2001 and 2002 it contained documentaries. Another risk is the topics and corpus drifting from being representative of real user communities and their tasks, which the TRECVID organizers hope is addressed by continually soliciting broad researcher and consumer involvement in topic and corpus definitions. An area that so far has remained outside of TRECVID evaluation has been the exploratory browsing interface capabilities supported by video collages and other information visualization techniques, which merits a final word in the next section.

7. DIFFICULTIES IN EVALUATING INFORMATION VISUALIZATION INTERFACES

It is tempting to introduce information visualization interfaces and admire them for their innovation without the benefit of empirical evaluation. Partly this is due to the difficulty of evaluating their complex interfaces. If low-level simple tasks are used for evaluation, it is easier to attribute differences in task performance to the different visualization attributes, but the simple tasks may bear little resemblance to real-world tasks. If complex tasks that come closer to real-world tasks are used, then more factors may confound the observed outcomes³¹. Another difficulty is in

determining the appropriate metrics to use. Measures for efficiency, satisfaction, and effectiveness are recommended in general¹, but these may be difficult to assess for visualization interfaces where browsing, querying, navigating, and scanning are all actions interwoven in the information access process^{32, 33, 34}. For example, do users who spend more time with a visualization system act so because it promotes exploration of potentially relevant areas, or are they spending more time because of problems comprehending the interface? For simple fact-finding tasks, effectiveness can be easily assessed, but the task is not well suited for visualization. If the user is asked to solve a precise information need, then the statement of that need can obviate the use of a browsing, exploratory interface (hence, the reason why exploratory visualization interfaces are not necessary for TRECVID tasks where the topics are stated with great text and visual detail). The user could just enter that precise query itself into the system and check the top answers. However, if the information need is more ambiguous and vague, then evaluation of effectiveness becomes tricky: was the need solved and to what degree? Good information visualization promotes a cycle of exploration and understanding that does not fit the traditional usability evaluation metrics of effectiveness and efficiency.

There is not enough space remaining in this paper to adequately address the issues of evaluating information visualization interfaces, but a recent article suggests three “first steps” to improving such evaluation³⁴: “the development of repositories of data and tasks, the gathering of case studies and success stories, and the strengthening of the role of toolkits.” The first two are within the realm of new directions for TRECVID: to help in the evaluation of information visualization interfaces targeting large video corpora by providing a test repository and suitable exploratory tasks, where such tasks are motivated and defined based on gathered case studies of real-world exploratory video use.

As a final example from my own work, consider the task of a student investigating “Colin Powell” using a CNN news corpus spanning 2001 to 2005. The Informedia collages allow the user to set views of interest, and Figure 5 shows two views active: face shots that are not anchor people, and a map view. The views update as the user manipulates a dynamic query slider on date, so that as the active date range varies from 2002 to 2003 to 2004 to 2005, the user sees the presentation in the upper left, upper right, lower left and lower right respectively. The interface is directly manipulable and allows immediate insights such as a set of people Powell is known to have met with in 2003 and 2004, the constant mention of Iraq and Afghanistan in stories dealing with Powell (even in 2005 as Powell exits the limelight), and the changing landscape of European countries in Powell stories through the years. An interesting future direction for TRECVID will be to initiate tasks and corpora allowing for the evaluation of such exploratory interfaces as these.

8. REFLECTIONS ON INFORMEDIA USER STUDY WORK AND FUTURE DIRECTIONS

Efficiency, effectiveness, and satisfaction are three important HCI metrics, with overlooking any of them reducing the impact of the user study¹. A mix of qualitative (observation, think-aloud protocols) and quantitative (transaction logs, click stream analysis, task times) metrics are useful, with quantitative data confirming observations and qualitative data helping to explain why. Discount usability techniques definitely offer benefit, to both iteratively improve video retrieval interfaces before committing to a more formal empirical study, and also to confirm that changes put in place as a result of a study had their intended effects. HCI guidelines can provide a jump start to the design process, as potentially useful automated video processing can be rendered inaccessible or useless through poor interface implementation²⁸.

The Informedia user study work has helped to direct some of the automated video processing approaches used by the Informedia research group as a whole, and endeavors to leverage the intelligence of the user to compensate for deficiencies in automated content-based indexing. Limitations include an over-reliance on accessible student and university staff populations as representative users, rather than fielding systems with other demographic groups. Other limitations include short term rather than longitudinal studies, and studies directed toward retrieval performance more than how well interfaces facilitate effective browsing. Informedia user studies have focused on particular video genres (news, documentaries) that have been collected, in part because they either came from willing Informedia partners or their intellectual property restrictions were not oppressive. The caution published with the studies and repeated again here is against generalizing results from such studies too broadly to other genres: video surrogates that work well for one genre may not be suitable for a different type of video³⁵. Goals for future Informedia work with greater impact include addressing these shortcomings by dealing with additional genres, experimenting with browsing tasks in addition to retrieval, and pursuing longitudinal studies with populations other than just university students.



Figure 5. Example of face and map views for Colin Powell query as date slider changes from 2002 to 2005.

As for the future of video summarization and browsing, I see at least three technical challenges. First, how can we address the semantic gap between low-level features and high-level user information needs for video retrieval, especially when the corpus is not well structured and does not contain narration audio documenting its visual contents? For example, what sorts of interfaces are needed to allow patients, families, nurses, doctors, and/or pharmaceutical companies to have access to a continuously recorded nursing home environment, where no narration has been added to

the video? As Alex Hauptmann notes in his CIVR keynote talk³⁶, and echoed by others³⁷, video retrieval researchers have successfully harvested low-hanging fruit: clever tricks and techniques of using speech transcripts and broadcast genres with detailed well-understood structures to identify the contents of news and sports broadcasts. The challenge now is to transform these techniques “into a serious body of science applicable to large-scale video analysis and retrieval”³⁶. Some directions include inferring media content from spatio-temporal context and the social community of media capture³⁸, the Large Scale Concept Ontology for Multimedia (LSCOM) work to reliably detect hundreds of intermediate semantic concepts like face, people, sky, and buildings across corpora³⁶, and working with less structured collections rather than just news or sports³⁷.

A second key problem for video retrieval is demonstrating that techniques from the computer vision community scale to materials outside of the researchers’ particular test sets, and that information visualization techniques apply more generally beyond a tested experimental task. I strongly believe in the value of community benchmarking activities like NIST TRECVID which support the statement from the 2003 ACM retreat report that “repeatable experiments using published benchmarks are required for the field to progress”³⁹, and would like to see TRECVID address browsing and information visualization evaluation more in the future.

Third, how can we best leverage the intelligence and goals of human users in accessing video contents meeting their needs, rather than overwhelming them with exponentially expanding amounts of irrelevant materials? Directions include applying lessons from the human computer interaction and information visualization fields, and being pulled by user-driven requirements rather than just pushing technology-driven solutions.

ACKNOWLEDGEMENTS

This work is supported by the National Science Foundation under Grant No. IIS-0205219 and by the Advanced Research and Development Activity under contract numbers H98230-04-C-0406 and NBCHC040037. CNN and others’ video contributions are gratefully acknowledged. Details about Informedia research, video contributors, and the full project team can be found at www.informedia.cs.cmu.edu. The user studies overviewed here were made possible by CMU HCII graduate students David Winkler, Adrienne Warmack, Neema Moraveji, and Ronald Conescu.

REFERENCES

1. Frøkjær, E., Hertzum, M., and Hornbæk, K. Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? *Proc. ACM CHI '00* (The Hague Netherlands, April 2000), 345-352.
2. Foraker Design. “Usability in Website and Software Design,” <http://www.usabilityfirst.com/>. Accessed Oct. 2005.
3. Wactlar, H., Christel, M., Gong, Y., and Hauptmann, A. Lessons Learned from the Creation and Deployment of a Terabyte Digital Video Library. *IEEE Computer* 32(1999), 66-73.
4. Christel, M., and Martin, D. Information Visualization within a Digital Video Library. *Journal of Intelligent Information Systems* 11(1998), 235-257.
5. Christel, M. Accessing News Libraries through Dynamic Information Extraction, Summarization, and Visualization. *Visual Interfaces to Digital Libraries LNCS 2539*, K. Börner and C. Chen, Eds. Berlin: Springer-Verlag, 2002, 98-115.
6. Christel, M., Winkler, D., and Taylor, C.R. Improving Access to a Digital Video Library. *Human-Computer Interaction: INTERACT97*, Chapman and Hall, London, 1997, 524-531.
7. Cox, R.V., et al. Applications of Multimedia Processing to Communications. *Proc. of IEEE* 86(5) (May 1998), 754-824.
8. Lienhart, R., Pfeiffer, S., and Effelsberg, W. Video Abstracting. *Comm. ACM* 40(12), 1997, 54-62.
9. Ponceleon, D., Srinivasan, S., Amir, A., Petkovic, D., and Diklic, D. Key to Effective Video Retrieval: Effective Cataloging and Browsing. *Proc. ACM Multimedia* (Bristol, UK, Sept. 1998), 99-107.
10. Yeo, B.-L., and Yeung, M.M. Retrieving and Visualizing Video. *Comm. ACM* 40(12), 1997, 43-52.
11. Zhang, H.J., et al. Video parsing and browsing using compressed data. *Multimedia Tools and Applications* 1(1) (1995), 89-111.
12. Kraaij, W., Smeaton, A.F., Over, P., and Arlandis, J. *TRECVID 2004 Proceedings*, <http://www-nlpir.nist.gov/projects/trecvid/>.

13. Ding, W., et al. Multimodal Surrogates for Video Browsing. *Proc. ACM Digital Lib.* (Berkeley, CA, Aug. 1999), 85-93.
14. Lee, H. and Smeaton, A.F. Designing the User Interface for the Físchlár Digital Video Library, *J. Digital Info.* 2(4), <http://jodi.ecs.soton.ac.uk/Articles/v02/i04/Lee/>, May 2002.
15. Nielsen, J., and Molich, R. Heuristic Evaluation of User Interfaces. *Proc. ACM CHI* (Seattle, WA, April 1990), 249-256.
16. Nielsen, J. Heuristic Evaluation. In Nielsen, J., and Mack, R.L. (eds.), *Usability Inspection Methods*. John Wiley and Sons, New York, NY, 1994.
17. Nielsen, J. Evaluating the Thinking Aloud Technique for Use by Computer Scientists. In Hartson, H. R. and Hix, D. (eds.), *Advances in Human-Computer Interaction* Vol. 3. Ablex, Norwood, NJ, 1992, 75-88.
18. Nielsen, J., Clemmensen, T., and Yssing, C. Getting Access to What Goes on in People's Heads? Reflections on the Think-Aloud Technique. *Proc. ACM Nordic CHI* (Aarhus, Denmark, Oct. 2002), 101-110.
19. Christel, M., and Moraveji, N. Finding the Right Shots: Assessing Usability and Performance of a Digital Video Library Interface. *Proc. ACM Multimedia* (New York, NY, October 2004), 732-739.
20. Boreczky, J., Girgensohn, A., Golovchinsky, G., and Uchihashi, S. An Interactive Comic Book Presentation for Exploring Video. *Proc. ACM CHI* (The Hague Netherlands, April 2000), 185-192.
21. Uchihashi, S., Foote, J., Girgensohn, A., and Boreczky, J. Video Manga: Generating Semantically Meaningful Video Summaries. *Proc. ACM Multimedia* (Orlando, FL, Oct. 1999), 383-392.
22. Christel, M., et al. Evolving Video Skims into Useful Multimedia Abstractions. *Proc. ACM CHI* (Los Angeles, CA, April 1998), 171-178.
23. Christel, M.G. and Warmack, A.S. The Effect of Text in Storyboards for Video Navigation. *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Salt Lake City, UT, May 2001), Vol. III, 1409-1412.
24. Christel, M.G., Hauptmann, A.G., Wactlar, H.D., and Ng, T.D. Collages as Dynamic Summaries for News Video. *Proc. ACM Multimedia* (Juan-les-Pins, France, December 1-6, 2002), 561-569.
25. Ahlberg, C. and Shneiderman, B. Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. *Proc. ACM CHI* (Boston MA, April 1994), 313-317.
26. Enser, P.G.B. and Sandom, C.J. Retrieval of Archival Moving Imagery - CBIR Outside the Frame? *Proc. Conf. Image and Video Retrieval (CIVR 2002)*, 206-214.
27. Shatford, S. Analyzing the Subject of a Picture: A Theoretical Approach. *Cataloguing and Classification Q.*, 6, 3 (1986), 39-62.
28. Christel, M., and Conescu, R. Addressing the Challenge of Visual Information Access from Digital Image and Video Libraries. *Proc. ACM/IEEE-CS Joint Conference on Digital Libraries* (Denver, CO, June 2005), 69-78.
29. Hauptmann, A., and Christel, M. Successful Approaches in the TREC Video Retrieval Evaluations. *Proc. ACM Multimedia* (New York, NY, October 2004), 668-675.
30. Christel, M., and Hauptmann, A. The Use and Utility of High-Level Semantic Features. *Proc. CIVR* (Singapore, July 2005), *LNCS 3568*: 134-144.
31. Kobza, A. An Empirical Comparison of Three Commercial Information Visualization Systems. *Proc. IEEE InfoVis* (San Diego CA, Oct 2001), 123-130.
32. Hearst, M.A. User Interfaces and Visualization. In Baeza-Yates, R., and Ribeiro-Neto, B. (eds.), *Modern Information Retrieval*, Addison Wesley/ACM Press, New York, 1999.
33. Hearst, M., et al. Finding the flow in web site search. *Comm. ACM* 45(9), 2002, 42-49.
34. Plaisant, C. The Challenge of Information Visualization Evaluation. *Proc. ACM Advanced Visual Interfaces* (Gallipoli, Italy, May 2004), 109-116.
35. Li, F., Gupta, A., Sanocki, E., He, L., and Rui, Y. Browsing Digital Video. *Proc. ACM CHI* (The Hague, Netherlands, April 2000), 169-176.
36. Hauptmann, A.G. Lessons for the Future from a Decade of Informedia Video Analysis Research. *Proc. CIVR* (Singapore, July 2005), *LNCS 3568*: 1-10.
37. Hart, P.E., Piersol, K., & Hull, J.J. Refocusing Multimedia Research on Short Clips. *IEEE Magazine* 12(3): 8-13.
38. Davis, M., King, S., Good, N., and Sarvas, R. From Context to Content: Leveraging Context to Infer Media Metadata. *Proc. ACM Multimedia* (New York, NY, Oct. 2004), 188-195.
39. Rowe, L.A. and Jain, R., *ACM SIGMM Retreat Report on Future Directions in Multimedia Research*, <http://www.sigmm.org/Events/reports/retreat03/sigmm-retreat03-final.pdf>, March, 2004.