Finding the Right Shots: Assessing Usability and Performance of a Digital Video Library Interface

Michael Christel
Computer Science Dept. and HCI Institute
Carnegie Mellon University
Pittsburgh, PA, USA 15213
1-412-268-7799
christel@cs.cmu.edu

Neema Moraveji
HCI Institute
Carnegie Mellon University
Pittsburgh, PA, USA 15213
1-412-268-7003
neema@cmu.edu

ABSTRACT
The authors developed a system in which visually dense displays of thumbnail imagery in storyboard views are used for shot-based video retrieval. The views allow for effective retrieval, as evidenced by the success achieved by expert users with the system in interactive query for NIST TRECVID 2002 and 2003. This paper demonstrates that novice users also achieve comparatively high retrieval performance with these views using the TRECVID 2003 benchmarks. Through an analysis of the user interaction logs, heuristic evaluation, and think-aloud protocol, the usability of the video information retrieval system is appraised with respect to shot-based retrieval. Design implications are presented based on these TRECVID usability evaluations regarding efficient, effective information retrieval interfaces to locate visual information from video corpora.

Categories and Subject Descriptors
H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems – *video.*

General Terms
Experimentation, Human Factors.

Keywords
Video retrieval, storyboard, TRECVID.

1. INTRODUCTION
Usability is the extent to which a computer system enables users to achieve specified goals in a given context of use effectively and efficiently while promoting feelings of satisfaction. Usability evaluation consists of methodologies for measuring the usability aspects of a system's user interface and identifying specific problems [2]. This paper reports on the usability evaluation of a storyboard-based video retrieval interface with demonstrated success in the TRECVID interactive retrieval tasks, success achieved when the system is operated by an expert familiar with the interface [1]. For the first time, performance metrics are gathered for novice users unfamiliar with the system. That experiment is reported here, and the novice users' interaction logs are considered along with heuristic evaluation and think-aloud protocol to identify usability problems and suggest improvements for shot-level video information retrieval systems.

A recent ACM strategic retreat examining the future of multimedia research identified three grand challenges, one of which is to "make capturing, storing, finding, and using digital media an everyday occurrence in our computing environment" [14]. The retreat report notes that with the widespread adoption of digital cameras and emergence of cell phones with built-in video cameras, coupled with increases in storage capacity and reductions in cost, we can now store massive amounts of image and video data, with the challenge being to make that data useful. This paper addresses the challenge by presenting an empirical report on video retrieval experiments and successes in the context of TRECVID, where TRECVID is noted in the same retreat report as valuable for the multimedia research community because it allows for repeatable experiments using published benchmarks. After an introduction to TRECVID and the video retrieval interface, results of the performance evaluation with novices are presented along with heuristic evaluation and think-aloud protocol. The paper concludes with design implications for shot-based retrieval from visually rich content sources like documentaries and broadcast news.

2. TRECVID AS EVALUATION FORUM
The National Institute of Standards and Technology has sponsored the Text REtrieval Conference (TREC) since 1992 as a means of encouraging research in information retrieval from large test collections. In 2001, the TREC Video Track (TRECVID) began with the goal to promote progress in content-based retrieval from digital video via open, metrics-based evaluation. Within TRECVID, the focus is on the shot as the unit of information retrieval, rather than the scene or story segment. Digital video has been decomposed into shots by a number of research projects and commercial systems in the past (see [6] for review), and shot boundary detection remains one of the tasks addressed by TRECVID [12]. A number of tasks ranging from semantic feature extraction to information retrieval are studied in the TRECVID forum, with this paper discussing interactive search involving a person in control of the retrieval process.

Interpreting search results for TRECVID 2001 proved difficult because of the lack of a shared shot reference defining a common

MM'04, October 10–16, 2004, New York, New York, USA.

unit for information retrieval. Since then, TRECVID publishes the delineation of the search test collection into a common shot reference, which is used by all participants in the query tasks and allows for easier assessment.

For TRECVID 2002, NIST defined 25 information needs to search for within a search test collection of 40.12 hours of MPEG-1 video from the Prelinger Archives and the Open Video archives [16]. The material consisted of advertising, educational, and amateur films produced between the 1910s and 1970s, spanning a wide quality and attribute spectrum, e.g., silent films and animated cartoons were part of the corpus. The search test collection was delineated into 14,524 shots for the common shot reference. For TRECVID 2003, NIST defined 24 interactive search topics for within a search test collection of 55.91 hours of MPEG-1 video: ABC World News and CNN Headline News from 1998, along with six hours of C-SPAN programming from mostly 2001. The search test collection was delineated into 32,318 shots.

The TRECVID interactive task is defined as follows: given a multimedia statement of information need (topic) and the common shot reference, return a ranked list of up to N shots from the reference which best satisfy the need, with the user having no prior knowledge of the search test collection or topics: N=100 for 2002, N=1000 for 2003. The topics reflect many of the sorts of queries real users pose, including requests for specific items or people, specific facts, instances of categories, and instances of activities [15, 16]. Mean average precision is used to compare the relative merits of the interactive systems.

In 2002, a system presenting image-rich interfaces, *System S*, achieved the best interactive search performance across 13 submitted interactive runs. In 2003, an improved version, *System S′*, produced the best performance, with users of *System S* producing the second highest performance across the 37 submitted interactive runs for the TRECVID 2003 evaluation. These results were obtained when the systems were driven by expert users, people who have been working with the research group that developed the systems for at least a year. Results from these system runs and their use by experts are discussed elsewhere [1]. An open question was whether the high-scoring interactive search achievements were due to the system itself, the knowledge of the experts, or some creative use of the system when in the hands of experts. In the hands of novices, would *System S′* still outperform the other interactive runs, i.e., how would it graph against the results shown in Figure 1?

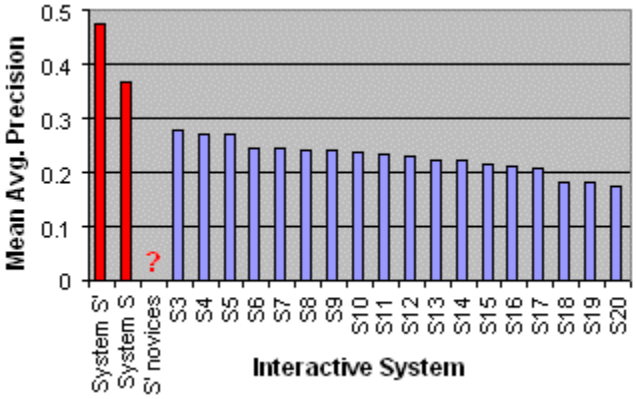

Figure 1. Mean average precision for top 20 TRECVID 2003 interactive runs: how will S' use by novices compare?

Other questions for investigation include whether the system would be used differently by novices, and what the novices' experiences tell us regarding image and video search and retrieval. The work reported here was motivated by three goals:

- Examine the information retrieval performance and experiences of novice users with *System S′* when applied toward the TRECVID 2003 tasks.
- Conduct discount usability techniques, namely heuristic analysis and think-aloud protocol, to clarify the benefits and deficiencies in the current interface.
- Inform the design and development of subsequent video information systems providing efficient, effective shot-based retrieval by both novice and expert users.

3. FEATURES OF SYSTEM S′

Both *System S* and *S′* work from the same automatic video processing. Through TRECVID, all participants have access to LIMSI-CNRS automatic speech recognized (ASR) transcripts [4], which coupled with closed-captions, provide some text metadata describing the broadcast news video [15]. We add our own speech recognition to this text metadata, used to provide another ASR transcript and also to time-align closed-captioning to specific seconds of video. In addition, we run text detection algorithms to find text in the broadcast, and pass that text first through visual filters and then off-the-shelf OCR systems to recognize the text and convert it into additional searchable descriptors for the video. This text metadata is coupled with cues like black frames and silences and used to segment sequences of video shots into stories, and the video stories are indexed using traditional term frequency/inverse document frequency text search statistics.

Both *System S* and *S′* emphasize storyboard layouts. A storyboard efficiently conveys video content by simultaneously displaying a series of thumbnail images each representing a video shot. Storyboards have been incorporated into numerous digital video retrieval systems [5, 6, 13, 14], with an acknowledged utility for navigating informational video-centric material [7]. The storyboard interface maintains temporal layout, accommodates contextual cues and filtering, provides a means of drilling down to synchronized points in the associated video, and presents a dense visual display of information back to the user. A storyboard resulting from a query on "Yasser Arafat" with the TRECVID 2003 corpus, is shown in part in Figure 2.

Figure 2. Top of video storyboard in System S′.

General problems with storyboard displays include the existence of too many shots in the represented set, and lack of screen real estate to show all the imagery. With *S* and *S′* the user had control over the size of the storyboard thumbnail size, from full MPEG-1 resolution 352 by 240 pixels, to half resolution in each dimension (176 x 120) to quarter (88 x 60) to the one-eighth (44 x 30) shown in Figure 2. One new feature introduced in *System S′* in 2003 was the addition of a magnifier tool feature under user control, whereby many tiny thumbnails could be browsed in a small screen area as in Figure 2 with the user able to show a full-resolution view of the thumbnail under the mouse with the press of the shift key, leading to a display as shown in Figure 3.

Figure 3. Anchored magnifier tool tip showing full-resolution details for shot at cursor position in storyboard.

Another means of reducing storyboard complexity is by showing only the matching shots, e.g., following text searches showing only the shots containing matching narrative or OCR text for the given query. For example, the Yasser Arafat query matched 22 segments consisting of 1351 shots, but the storyboard can be reduced to just 36 matching shots (the first 24 are shown in Figure 2), shots during which "Yasser" and/or "Arafat" is spoken or appears as text on the visual display.

Information in documentaries and news is conveyed in both the audio and video [6, 7]. For some topics, like celebrities and politicians, the narrative mentions names and the reasons why individuals are in the news, with the visuals showing the people. Here, ASR and overlay text are ideal indexing strategies to locate relevant shots and display them in storyboards. For peripheral topics such as generic people in the street or crowded road traffic, the narrative does not describe the setting adequately to precisely find such material. For those cases, the visual channel holds most or all of the information. Similarly, material more likely to be found in commercials (advertisements) rather than news, such as cups of coffee or cats, will not be described by the commercial's audio narrative, as that is devoted to selling a product rather than describing the visual scene.

In addition to text query in both *S* and *S′*, image query based on color or texture is another means by which the user can explore the corpus. A thumbnail can be dragged into the query window to initiate an image-based search using that image as a search key. A third means of visual browsing, one newly introduced in *System S′,* makes use of the TRECVID image feature classes [15] to support investigations into visually oriented topics. Through automatic analysis, shots are classified with TRECVID features including animal, building, cityscape, people, and road, as well as a few genre-specific features appropriate for news such as weather, news subject face, and commercial; TRECVID classifier evaluations show typical feature mean average precision of less than 0.4. The top-scoring few hundred shots are precomputed into storyboards which the user can selectively load, e.g., in response to a topic requiring urban scenes the user can load the top-scoring cityscape shots.

Within *S* and *S′*, storyboard views are produced by text or image query, or browsing precomputed feature classes (for *S′* only). Another view shows a single thumbnail shot per segment following the query, with a button to open all of the shots for that segment in yet another storyboard view. Video is played starting at the shot represented by the thumbnail by double-clicking on it, with a right-click action marking the shot as part of the answer set for a given topic. The TRECVID topics are described with both text and multimedia examples: the 24 interactive search topics are listed in Table 1. Targets are shots visually showing the topic, e.g., a shot of Arafat rather than a news person talking about Arafat. For the search runs overviewed in Figure 1 and conducted with novices as reported in Section 4, users have a time limit of 15 minutes per topic. The answer set is a unique storyboard view in that the user's actions populate it and the thumbnails can be reordered and deleted. Further discussion of the automatic processing leveraged by the interfaces is published elsewhere, along with an overview of the use of the systems by experts [1].

Table 1. TRECVID 2003 interactive search topics.

Type	Generic	Specific
Find objects	Cat Cup of coffee Helicopter Tank	Sphinx Tomb of Unknown Soldier Mercedes logo
Find people	Person diving Urban people	Osama bin Laden, Morgan Freeman, Pope John Paul II, Yasser Arafat, Mark Souder
Find events	Rocket launch Airplane take-off Baseball pitch Incoming train Basketball hoop	
Find scenes	Fire, Aerial views, Road with cars, Snowy mountain	White House fountain

While the interface has had demonstrated success with expert users, we wanted to determine the performance and usability of the system when operated by novices who had never seen the system before and were unfamiliar with TRECVID and shot-based video retrieval. We collected usability metrics through direct user testing and inspection techniques. Specifically, we observed users interacting with the interface through performance logs and think-aloud sessions, and also had evaluators use a set of

criteria or heuristics to identify potential usability problems in the interface.

4. SEARCH PERFORMANCE

Twelve undergraduate and graduate students (age range 18-29, 7 female) were recruited at Carnegie Mellon University to participate in a 90 minute study. Each student worked independently on an Intel® Pentium® 4 class machine, with a 1600 x 1200 pixel 21-inch color monitor. They received a 15-minute interactive multimedia tutorial introducing *System S′* and then used the system to answer four TRECVID 2003 topics, limited to 15 minutes each. Students were paid $15 for their efforts. As with most of the top 20 runs shown in Figure 1, multiple users combined to complete the 24 TRECVID topics. Each run includes the results for all 24 topics, and each topic's results come from a single user, but a user may answer from 1 to 24 topics. In our experiment, we produced two runs of six subjects each. Specifically, the first subject completed topics 1 through 4, the next topics 5 through 8, etc., so that 6 subjects combined to complete a full interactive run through the 24 topics.

Using the same answer key as was used for Figure 1, the first six subjects produced a mean average precision score of 0.383. The second six subjects produced a mean average precision score of 0.381. The scores are remarkable in that they are very close, increasing our confidence in placing the performance of *System S′* with novices second on Figure 1's ranking, below only the performance of *System S′* with the expert (0.476) and above the use of *System S* by experts (0.368).

These scores represent the most conservative comparison, using the answer key from NIST prior to the TRECVID 2003 workshop. However, the answer key from NIST assessors was generated by looking at the top 50 (or more) shots submitted per topic per interactive run and manually grading the shot as correct or incorrect. The subjects produced 131 shots, from their top 50 per topic, which were not graded. These were conservatively counted as incorrect in the above paragraph's statistics. In accordance with the TREC pooled grading procedure, we then graded these shots independently by two assessors. The interrater reliability was 93.1%; the 9 shots with ambiguous grading were assessed by a third person to make the final decision. 41 previously unidentified correct shots across the topics were found, and when the answer keys are updated with these shots, the mean average precision for the first six subjects increases to 0.390, and for the next six to 0.391. The mean average precision score for the *System S′ with expert* run drops slightly to 0.472, and the *System S* run to 0.367. Using the more accurate comparison with the revised answer key, the use of *S′* by novices firmly establishes itself as the second-best interactive run, behind only the use of *S′* by the expert, as shown in Figure 4. We conclude that *S′* has interface features enabling effective shot-based retrieval from news video collections by both novices and expert users.

Figure 5 shows the precision-recall curves for the averaged two novice runs across the 24 topics, compared to the expert runs and the averages of the next-best runs of systems S3,S4,S5 and S6,S7,S8 shown in Figure 4. As information retrieval performance decreases, the curve continues to depress closer to the origin, so plotting runs S3, S4, etc. independently and graphing down to run S20 would produce a cluster of lines in the lower left of the plot at and below the dotted lines. This figure shows a more detailed view into the performance of the novices, who did comparatively well with respect to the expert using the same system *S′* while outperforming the experts that used the older system *S*, especially to the left of the curve reflecting the goal of high precision. Novice users did an excellent job with the *S′* interface finding correct shots, better than the experts using an older system.

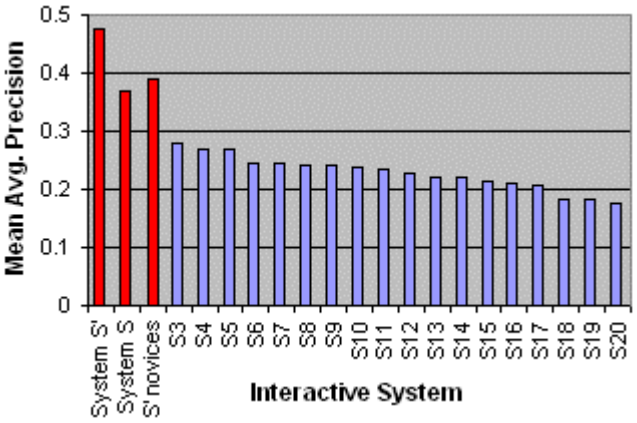

Figure 4. Mean average precision: novice runs second-best.

The performance differences between the systems diminish if high recall is the primary goal (the right portion of the curve), i.e., retrieving all of the correct shots for a topic. This outcome is expected because both *S* and *S′* were designed to allow for and encourage visual inspection and approval of candidate shots for the answer set. With some topics having hundreds of correct shot answers, and a time limit of 15 minutes per topic, visual inspection of all the correct answers becomes improbable, so recall suffers at the expense of precision. We made this design choice so that our interactive systems could employ the user's attention to maximize precision, while in parallel working on fully automatic retrieval systems with no user in the loop that focus on better indexing, search, and machine learning algorithms to accomplish high recall of large answer sets. Our ultimate goal is to combine these two efforts: automated retrieval will present large candidate sets with good recall to a user for inspection, who can then filter the set down quickly to achieve better precision.

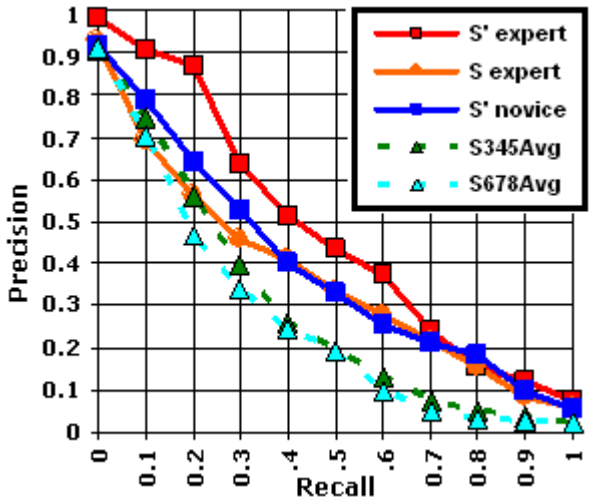

Figure 5. Precision-Recall across 2003 TRECVID 24 topics.

The performance metrics show that in the hands of novices, *System S′* performs better than an older system *S* operated by experts. The improvements made from *S* to *S′* allow both novices and experts to achieve high mean average precision scores on shot-based retrieval tasks. Through additional evaluation, we look at aspects of the interface contributing to the performance achievement and areas of the interface where further improvements are necessary.

An analysis of the interaction logs shows that novices relied on text query to answer topics. A few other summary statistics from the logs are shown in Table 2, with each novice user completing 4 topics, 15 minutes per topic. A column shows the results from the expert's run (24 topics, 15 minutes per topic) with *S′* in 2003.

The nature of the text queries was the same with expert and novices: all used brief, directed queries. However, the novice users relied much more on text search, and their queries produced larger sets. The expert user employed image queries and the precomputed sets, i.e., the best shots automatically determined to have a feature like "roads", while novices did so rarely. Novices avoided image search because it took too long. Novices did not use precomputed sets because their access was hidden in pull-down menus, and their utility in answering topics was not obvious without additional experience in their use.

Table 2. Summary statistics from novice user interaction logs.

Statistics Reported as Averages	Novice	Expert
Text queries issued per topic	10.0	2.5
Word count per text query	2.2	2.1
Number of video story segments returned by each text query on avg.	74.8	56.4
Image queries per topic	0.4	1.3
Precomputed feature sets (e.g., "roads") browsed per topic	0.5	1.3

5. THINK-ALOUD PROTOCOL

Three undergraduate students and one university employee (three males) were recruited at Carnegie Mellon to participate in a one hour think-aloud protocol using the TRECVID 2003 topics. These users were shown the same online tutorial introducing features of the system as was used in the performance evaluation, but here their work was not restricted to 15 minutes per topic, and they were asked to think aloud, i.e., describe the contents of their working memory, while they performed the task with someone observing them and taking notes. The users had no prior experience using the system. By listening to users think aloud while performing tasks, we could infer their line of thought, what they expected from the system, what problems they faced, and their reactions. Pioneering work with think-aloud protocol show it to be an effective "discount" usability engineering method, capable of uncovering errors at cost savings after trials with only a few users [10]. Its primary disadvantage is that requiring the user to reflect on and vocalize the interactions with the system introduces added burden on the user which would not be there in actual system use. The observer plays an interruptive role, the users face increased strain and cognitive load, and users may not vocalize when the task becomes difficult. A recent examination of think-alouds notes that the technique is typically employed along with other usability engineering methods to compensate for its deficiencies [8]. In our case, we ran performance evaluations with strict timing and no interruption of users (Section 4), and conducted four think-aloud sessions with novice users to augment the performance data with feedback as to what the novice user is thinking about and struggling with in the interface.

The results from the think-aloud sessions help explain problems experienced by users in conducting TRECVID tasks, providing insights into the interactions logged for the 12 users discussed in Section 4. The sessions also illustrate the symptoms experienced by novice users if certain usability heuristics are not followed (Section 6). The results in general are clustered and summarized in Section 7.

To illustrate the contribution of think-aloud protocol here, consider the following uncovered problem. Figure 2 shows a storyboard of matching shots following a query on "Yasser Arafat." Another window is also shown to the user following the query, a view with one thumbnail per video story or segment, allowing for easy access to the 22 story segments matching the query, shown in Figure 6. The rationale for developing the two views was to allow for easy access to shot level detail (Fig. 2) and to segment-level detail (Fig. 6). The user thinking, as shown by the protocol, was confusion over the two views. Both views show thumbnails at the same resolution. Repeatedly across different users and topics, the user did not understand that in the one case (Fig. 2), thumbnails represent shots, while in the other (Fig. 6), thumbnails represent segments.

Figure 6. Segment-based view of "Yasser Arafat" query shown simultaneously with matching shot view (Fig. 2).

6. HEURISTIC EVALUATION

Heuristic evaluation [9, 11] is a usability engineering method for finding the usability problems in a user interface design so that they can be attended to as part of an iterative design process. Heuristic evaluation involves having a set of evaluators independently examine the interface and judge its compliance with recognized usability principles (the "heuristics"). The findings are aggregated, producing a prioritized list of usability problems in the interface with references to principles that were violated by the design in each case in the opinion of the evaluator.

Three user interface experts conducted a heuristic evaluation on *System S′* using the data and tasks from TRECVID 2003, working from the ten usability heuristics published by Nielsen [11]. 56 error cases were found, with most errors classified against the "consistency and standards", "recognition rather than recall", and "aesthetic and minimalist design" heuristics. A sampling of errors

which can be understood by referring to Figures 2, 3, and 6 are presented in the next three paragraphs.

There is a mismatch between the terminology used in the text in the interface and what novice users understand. While ACM Multimedia Conference attendees understand and appreciate the hierarchy of video broadcast – segments – shots, the terms "shot", "segment", "storyboard", and "result" are not clearly defined in the interface. Their use in window captions and titles does not adequately clarify the separate views or allow the user to distinguish Figure 2's view from Figure 6's.

In terms of consistency, the thumbnail representation currently affords different operations depending on its parent view. Regardless of its placement, the thumbnail should allow a set of consistent operations, including letting the user play video at that point, initiate an image search, see a full-resolution rendering like Fig. 3, add the shot to the topic answer set, copy the image to the clipboard, etc. The add-to-answer-set operation that is hidden as a right mouse button click should be made more visible, with all the operations above listed in a pop-up context menu on right-click, in accordance with the operating system conventions where the system is hosted. Other hidden operations should also be made more visible, including the full-resolution rendering of Fig. 3, initiating an image search, and loading pre-built sets of shots automatically detected to have some feature like "roads."

Right-click produces a pop-up context menu over the video playback area and text areas, but not the thumbnail image area. Tool tip context-sensitive help pops up over some but not all button and selection controls. Long load times for search results or large image sets are not communicated properly through mouse hourglass cursors, status bar messages, progress bars, and other conventions.

Reflecting on the use of heuristic evaluation along with user testing evaluation, we reach the same conclusions as found in an earlier comparative evaluation of usability methods applied to a relational database retrieval system [3], namely, that "in order to fully assess an interface it is necessary to use a variety of techniques: there might be whole classes of error missed by any one." Direct user testing finds symptoms of problems, whereas heuristic testing focuses on identifying the cause of the problem. For example, user testing showed the confusion between the shot view of Figure 2 and segment view of Figure 6. Heuristic evaluation signalled one cause of the problem as a mismatch between the system and the world, i.e., between the use of terms "shot" and "segment" and their interpretation by novice users.

As noted in the earlier study [3], a failure to understand the underlying cause has implications for redesign as a new design may remove the original symptom, but if the underlying cause remains, a different symptom may be triggered. Heuristic evaluation provides causal categories to better analyze observed usability problems, but observation of novices is still vital as many problems are a consequence of the users' knowledge, or lack of it, when interacting with a system on actual tasks, in our case the TRECVID shot-based retrieval tasks.

7. DESIGN IMPLICATIONS

By running the TRECVID 2003 interactive search experiment with novice users, we have shown that the improved *System S′* produces high-scoring interactive retrieval performance, second only to an expert's use of the same interface. Direct user tests and inspection techniques reveal a number of fixable problems with the interface restricting the novices from achieving higher levels of performance. As we continue with iterative refinement and evaluation of the interface, we look to address the following problem areas for TRECVID 2004 shot-based video retrieval and beyond.

7.1 Capturing User Interaction History

In working through the TRECVID topics there are many judgments made by the user regarding shot relevance which are not leveraged. Specifically, the user may see the shots of Figure 2 and decide that the second, fifth, and sixth shots are relevant. In the current interface, these shots' thumbnails are rendered as grayscale images as soon as the user adds them to the answer set, to convey that they have already been added and need not be considered any further. Two problems remain. First, the grayscale rendering is too subtle, difficult to distinguish in scenes that lack vibrant colors. Second, the user may also pass judgment on shots that are not relevant, e.g., explicitly noting that the first, third and fourth shots in Figure 2 are not relevant. Such judgment may take place at the segment level of granularity too, e.g., the user may judge the second segment of Fig. 6 (and hence all of its shots) as irrelevant. A history of the shots that the user explicitly passes judgment on as being relevant or irrelevant can reduce the number of shots the user is shown in follow-up queries. If the user queries on "Palestine" and some of the judged shots in Figure 2 and 6 are part of the result set, these already-judged shots can be suppressed in favor of only showing shots yet to be judged.

Design decisions include how to efficiently mark that a user has judged a shot or segment as irrelevant. Placing the shot in the answer set can remain as the indicator for a shot being judged as relevant. Different means of marking already-judged shots, including other image effects besides grayscale, the use of border techniques, and the immediate hiding of judged shots, will be investigated. Inspecting judgment histories, modifying them, and clearing them are features that can be shielded from novice users but then progressively disclosed to migrate novices into expert users having great control over the interface presentation.

7.2 Providing Consistent Features

The mouse-over full-resolution image zoom feature shown in Figure 3 was used heavily and appreciated by users. However, it was not available consistently across all displays of thumbnails, nor were other thumbnail-based features like playing the video corresponding to that thumbnail's shot and initiating an image search from the thumbnail. Consistency will be emphasized when updating *S′*, including consistency of image/thumbnail based operations to video/player based operations. For example, a right click on the thumbnail adds it to the answer set for a topic. Users expected that a right click in the video player area would add the shot currently being played into the answer set as well.

Where there are differences in representation that argue against consistency, we will prototype potential solutions and run follow-up usability tests on the pilot interfaces to determine the better design. For example, the thumbnails in Figure 6 represent segments, so double-clicking on one plays video for the chosen segment starting at the beginning of the segment, while double-clicking on a thumbnail in Figure 2 plays its segment but starting at the beginning of the shot represented by the thumbnail, since in that view the thumbnail represents a shot. In this testing iteration,

novices requested the operation of a "quick preview" similar to Figure 3 except the thumbnail's shot video is played from start of shot to end of shot rather than just showing the full-resolution image. A question arising from adding a "quick preview" to thumbnails is what to do with respect to the segment-based view of Figure 6. If the user initiates a quick preview on a thumbnail there, is the expectation that just the one shot's video will be previewed, or that the segment's full video is previewed?

7.3 Enabling Informed Choices

The segment view of Figure 6 is ranked in order of relevance by the search engine returning the results. This ranking helped users with their decisions: higher-ranked segments were more likely to contain relevant shots. The text titles that progressively displayed as the user moved the mouse cursor over the results also helped: these titles are automatically generated from the ASR transcripts and other text metadata for the video. However, the thumbnail representing the shot had a strong influence over the user's investigative pattern: a visually interesting thumbnail for a segment low in the relevance ranking was more likely to be selected for inspection (playing the video and/or showing the storyboard of shots for that segment) than an unappealing thumbnail for a segment high in the relevance ranking. Domain-specific filtering of the choice of thumbnail for the segment can help by removing the unappealing candidates masking potentially relevant segments. Specifically, we have implemented an improved anchorperson detector to greatly reduce the chances of segments being represented by an anchor person (like segments 9 and 10 in Figure 6). Other filters for news include reducing weather and commercial shots.

Another option is to show more visual information per segment in the view of Figure 6, e.g., represent segments with more than one thumbnail or allow progressive disclosure of additional imagery detail, just as the additional text title detail is shown when the user focuses on a segment. Taken to the extreme – showing all the shots for a segment that match a given query – leads to the interface shown in Figure 2. The correspondence between the matching shots following a query (e.g., Fig. 2) and the matching segments following a query (e.g., Fig. 6) needs to be better communicated to the novice user through better terminology, layout, and visual cues. Multiple views are useful, as indicated by the logs from the 12 novices performing timed topics. 58% of their answer shots were taken from the matching shots interface (as shown in Fig. 2), with the rest of the answer shots taken from segment-based views as shown in Fig. 6 or from a storyboard for a single segment launched from such segment views.

7.4 Facilitating Efficient Investigations

There were three primary means of navigation with the tested *System S′*: text query, image query, and browsing by top ranked visual features. Text search performed quickly, returning within a few seconds with results and representative thumbnails. Image search took on the order of twenty seconds to complete, and after being used at most once or twice was abandoned more because of its slow retrieval time than for any mismatch between user's expectations of content-based retrieval and the actuality of color-based or texture-based retrieval. Browsing by features loaded the precomputed shot sets quickly, but was an option hidden in pull-down menus that were not accessed much. The users also questioned the believability and utility of the shot sets based on their low precision, e.g., the "roads" shot set holds roughly 40% road shots in its set of 400.

Future work includes making the current navigation mechanisms more efficient, especially image searching, through better indices, database improvements, and caching architectures. Browsing feature sets may be improved by restricting inclusion in the sets even further to only the highest ranked shots for that given feature, and continued improvements on automatic feature classification. Advanced search strategies need to be tucked away from initial view but available through an easily seen interface control like an "Advanced…" button on the query window. When selected, the advanced search can note how to accomplish and/or searches, phrase searches, and searches within specific fields like just the video OCR text or just the ASR text. HCI texts argue for such scaffolding of the interface, so that an easy-to-use system is accessible immediately by novices, with advanced features available for use as the novices gain experience so that they can gradually become expert users with greater control to achieve high levels of performance.

8. FUTURE WORK

We focused primarily on an interaction metaphor of "search – examine results – search again" without much investigation into broad browsing strategies for visual content. The storyboard views have been used as communicators for results of queries, to browse within the query result sets. They can be employed as a way to browse the full corpus as well, with of course the problem being one of scale. Browsing 32,318 images for TRECVID 2003 might be possible with some manner of clustering and organizing into visual hierarchies, but will such approaches still work for browsing a few thousand hours of news broadcasts containing 2 million shots? Research directed toward large digital photograph collections may also apply to solutions for using storyboards as up-front video browsing interfaces without the need of an initial query. An implicit goal of the TRECVID evaluations is to help chart the progress of automatic feature classifiers like face, people, outdoors, and cityscape, showing that perhaps these classifiers will reach the level of maturity needed for their use as efficient, effective complementary filters for storyboards. For example, in looking for shots of people walking in urban environments, the user could start, not with a text or image query, but a browse through a storyboard holding all the shots in the corpus rated highly as containing people, person activity, and cityscapes or buildings, assuming the precision for these classifiers improves enough to warrant their use.

The report from the recent ACM retreat examining the future of multimedia research noted that context could be used more thoroughly in multimedia interfaces [14]. The relative high information retrieval performance of *System S′* is due to its reliance on an intelligent user possessing excellent visual perception skills to compensate for comparatively low precision in automatically classifying the visual contents of video. The user sets the context through a query, and that query is used to reduce candidate shots from tens of thousands to tens or hundreds. The storyboard views facilitate quick browsing, with full-resolution detail or video playback on demand, and quick access to neighboring matching shots or all the shots in the segment.

Context can be much more than just query context, however. Further context can come from external information sources like

the Web and structured information sources like dictionaries and thesauri. These sources can be used to improve the recall performance of the system and the coverage of the query, e.g., if looking for a basketball going through the hoop, a query on "basketball" could be automatically expanded to include related terms like "three-point shot", "foul shot", "dunk", and names of all the NBA basketball teams to retrieve relevant stories that do not specifically mention "basketball." Semantic query expansion based on using related imagery from the Web as additional image search keys has also been explored. These strategies tend to vastly increase the candidate result set of shots, with context from genre and usage having the potential to reduce the shot set back to manageable size.

Context can be material-specific, coming from the genre of the video itself [7]. For news broadcasts, shots can be grouped into studio shots, anchor person shots, reporter shots, weather report shots, and commercial (advertisement) shots. A news broadcast can be decomposed into promotional segments (e.g., "Coming up next, …."), introductory title sequences, sign-off sequences ("This is Dave Smith reporting from Atlanta."), and other sorts of segment sequences specific to a particular producer's news format and style. Genre context can be used to filter an expanded shot set to just the likely candidates, e.g., eliminating anchors, reporters, weather reports and studio shots when looking for road traffic and cars.

Another type of context is the usage context from the user's interactive session. Earlier we discussed marking which shots have already been judged by the user as relevant or irrelevant for a given topic, and suppressing those shots from display in subsequent interactions regarding the topic. User interactions can also provide cues as to the common, typical features of shots marked as relevant. For example, if they tend to come from the same broadcaster in the same time period and show faces, other face shots from that broadcaster and time period will be prioritized in the sets returned from semantic expansion on follow-up queries. Mining the users' activity can improve the ordering and reduce the number of shots shown in subsequent interactions.

Our future work will explore the use of context in such ways to improve the recall performance of *System S'*. Through the use of performance evaluations using open testing procedures, metrics, and data, the benefits of future video information retrieval systems can be better assessed. Through the use of usability inspection techniques and direct user testing, an iterative design process can be supported for refining interfaces providing efficient, effective access to relevant shots from video collections.

9. ACKNOWLEDGMENTS

This material is based on work supported by the Advanced Research and Development Activity (ARDA) under contract numbers H98230-04-C-0406 and NBCHC040037. We thank the members of the Informedia research team who make this work possible (http://www.informedia.cs.cmu.edu), especially Chang Huang for her interface development and heuristic evaluation.

10. REFERENCES

[1] Christel, M., Huang, C., Moraveji, N., and Papernick, N. Exploiting Multiple Modalities for Interactive Video Retrieval. In *Proc. ICASSP 2004* (Montreal, Canada, May 2004), Vol. III, pp. 1032-1035.

[2] Dix, A., Finlay, J., Abowd, G. and Beale, R. *Human Computer Interaction*, 2nd Edition. Prentice Hall, 1998.

[3] Doubleday, A., Ryan, M., Springett, M. and Sutcliffe, A. A Comparison of Usability Techniques for Evaluating Design. In *Proc. ACM Conf. on Designing Interactive Systems* (Amsterdam, Aug. 1997), pp. 101-110.

[4] Gauvain, J.L., Lamel, L. and Adda, G. The LIMSI Broadcast News Transcription System. *Speech Communication*, 37, 1-2 (2002), pp. 89-108, ftp://tlp.limsi.fr/public/spcH4_limsi.ps.Z.

[5] Komlodi, A., and Marchionini, G. Key Frame Preview Techniques for Video Browsing. In *Proc. ACM Digital Libraries* (Pittsburgh, PA, June 1998), pp. 118-125.

[6] Lee, H. and Smeaton, A.F. Designing the User Interface for the Físchlár Digital Video Library. *J. Digital Info, 2*, 4 (May 2002), http://jodi.ecs.soton.ac.uk/Articles/v02/i04/Lee/.

[7] Li, G., Gupta, A., Sanocki, E., He, L. and Rui, Y. Browsing Digital Video. In *Proc. ACM CHI 2000* (The Hague, The Netherlands, April 2000), pp. 169-176.

[8] Nielsen, J., Clemmensen, T., and Yssing, C. Getting Access to What Goes on in People's Heads? Reflections on the Think-Aloud Technique. In *Proc. ACM Nordic CHI* (Aarhus, Denmark, Oct. 2002), pp. 101-110.

[9] Nielsen, J., and Molich, R. Heuristic Evaluation of User Interfaces. In *Proc. ACM CHI 1990* (Seattle, WA, April 1990), pp. 249-256.

[10] 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, pp. 75-88.

[11] Nielsen, J. Heuristic Evaluation. In Nielsen, J., and Mack, R.L. (eds.), *Usability Inspection Methods*. John Wiley & Sons, New York, NY, 1994.

[12] NIST, *Digital Video Retrieval at NIST: TREC Video Retrieval Evaluation*, 2001-2004, http://www-nlpir.nist.gov/projects/trecvid/.

[13] Ponceleon, D., Srinivasan, S., Amir, A., Petkovic, D., and Diklic, D. Key to Effective Video Retrieval: Effective Cataloging and Browsing. In *Proc. ACM Multimedia* (Bristol, UK, Sept. 1998), pp. 99-107.

[14] Rowe, L.A. and Jain, R. ACM SIGMM Retreat Report on Future Directions in Multimedia Research, March 2004, www.acm.org/sigmm/main/events/sigmm_retreat/sigmm-retreat03-final.htm.

[15] Smeaton, A., Kraaij, W., and Over P. The TREC Video Retrieval Evaluation (TRECVID): A Case Study and Status Report. In *Proc. RIAO 2004* (Avignon, France, April 2004).

[16] Smeaton, A.F., and Over, P. The TREC-2002 Video Track Report. In *TREC 2002 Proc.* (NIST, Gaithersburg, MD, Nov. 2002), http://trec.nist.gov/pubs/trec11/t11_proceedings.html.