Examining User Interactions with Video Retrieval Systems*

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 reports on TRECVID 2005 and 2006 interactive search tasks conducted with the Informedia system by users having no knowledge of Informedia or other video retrieval interfaces, but being experts in analyst activities. Think-aloud protocols, questionnaires, and interviews were also conducted with this user group to assess the contributions of various video summarization and browsing techniques with respect to broadcast news test corpora. Lessons learned from these user interactions are reported, with recommendations on both interface improvements for video retrieval systems and enhancing the ecological validity of video retrieval interface evaluations.

Keywords: user studies, TRECVID, information visualization, digital video library, video surrogate, user interface evaluation, video retrieval, Informedia, human-computer interaction

1. INTRODUCTION

A number of Informedia user studies have taken place through the years, most often with Carnegie Mellon students and staff as the participants. These studies were surveyed in a 2006 paper reporting on how they can provide a user pull complementing the technology push as automated video processing advances[1]. The merits of discount usability techniques for iterative improvement and evaluation were presented in that same survey paper, 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. Conclusions were 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. This paper complements that 2006 survey report by presenting a series of user studies conducted with representatives of a user community outside of the college/university population: professional situation analysts whose jobs focus on the management, analysis, processing, and dissemination of strategic and tactical intelligence from varied, typically voluminous data sources.

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 usability[2]. Before surveying the Informedia user studies conducted with the analysts, a discussion of ecological validity is warranted, because it affects the impact of the user study results. Foraker Design defines ecological validity as follows[3]:

> **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 this paper. In fact, the concern that the "users may not represent actual users" led to the interest in employing actual situation analysts in the interface assessments reported here, rather than conduct trials with college students. The conclusions regarding TRECVID interactive search tasks are interesting in that the TRECVID video retrieval research community, who often pose as users for these tasks against their own developed systems, respond much differently than do the "real users" as represented here by the situation analysts.

2. NIST TRECVID VIDEO RETRIEVAL EVALUATION

The NIST 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 from 12 to 69 companies and academic institutions from 2001 to 2006[4]. 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 (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 have been 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.

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[4, 5]. 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 pictures[6]. In video retrieval, a broadcast is commonly decomposed into numerous shots, with each shot represented by a keyframe: a single bitmap image extracted from that shot. The numerous keyframes can then be subjected to image retrieval strategies. This simplified approach to video retrieval is taken here, with the benefit that many lessons learned for such shot-based video retrieval will be applicable as well for digital still image (photograph) retrieval. Two shortcomings left for separate investigations are determining how to best represent the contents of a shot with a keyframe or set of keyframes, and how to account for and leverage temporal information present in video, such as rate and direction of motion and camera pans and zooms. In the Informedia interfaces discussed here, the temporal progression of information in video is presented through the arrangement of shot thumbnail imagery into storyboards, where shots related in video time within the same video segment are displayed sequentially in time order.

3. INTRODUCTION TO THE VIDEO RETRIEVAL STUDIES

This report details the testing procedures and results from an assessment of the Carnegie Mellon University (CMU) Informedia/ENVIE (Extensible News Video Information Exploitation) system conducted in September, 2006. The system, henceforth referred to as ENVIE, was produced through ARDA funding under both the AQUAINT and VACE programs as listed in the acknowledgments. Three different broadcast news data sets were used for the assessment, two of which were provided through the NIST TRECVID video retrieval evaluation forum[4]. The 85-hour TRECVID 2005 test set consists of 140 movies from CNN, NBC, MSNBC, and Arabic and Chinese broadcast news sources, covering the time period from November 16 to December 1, 2004. These movies were partitioned into 4393 story segments and 77979 video shots through ENVIE processing. The larger 166-hour TRECVID 2006 corpus consists of 259 movies from U.S., Arabic, and Chinese news sources covering November/December 2005. These movies were partitioned into 5923 segments and 146,328 video shots through ENVIE processing. The third corpus was collected by CMU from a video cable news feed and represents the most current, largest set used in the assessment: 240 hours of CNN and Chinese news sources from January to May, 2006. This 2006 news corpus was partitioned into 11191 segments and 183,654 shots. ENVIE was used over the course of two days by six analysts on both structured and exploratory information retrieval tasks, with the work breakdown given in Table 1. User study assessment techniques included a within-subjects formal experiment, timed controlled experiments, transaction logs of all keyboard and mouse interactions with the interface, think-aloud protocols, questionnaires, group and individual interviews. The questionnaires accompanying the TRECVID topics was the same as used across all of the TRECVID 2004 interactive search participants (26 international groups), designed based on prior work conducted as part of the TREC Interactive track for several years[7]. The merits of these various assessment techniques are discussed further below.

Table 1. Work breakdown (in the same order as performed by the participants) for ENVIE assessment.

Task per Participant	Corpus	Description
Practice topic, then 4 TRECVID 2005 topics	TRECVID 2005 test set	Within-subjects experiment, with first 2 topics in Interface A, second 2 in Interface B
4 TRECVID 2005 topics	TRECVID 2005 test set	Within-subjects experiment, with first 2 topics in Interface B, second 2 in Interface A
Watch demonstrations of ENVIE capabilities	2006 International News (CMU corpus)	Educate participants in info. visualization, summarization capabilities of ENVIE
Hands-on use of ENVIE on exploratory tasks, both seeded and self-nominated	2006 International News (CMU corpus)	Transaction log recording from exploratory search tasks; individual questionnaires and group interviews as well
4 TRECVID 2006 topics	TRECVID 2006 test set	Full-featured ENVIE system with questionnaires
Seeded exploratory task plus “talk through the effort”	2006 International News (CMU corpus)	Think-aloud protocol
Wrap-up reflections	All	Individual interviews

The six situation analysts (five male, one female), represent a particular consumer pool for news corpora: people mining open broadcast sources for information. These analysts, compared to the CMU students and staff for the prior reported studies surveyed in the 2006 report[1], were older (2 older than 40, 3 in their 30s), more familiar with TV news, just as experienced with web search systems and frequent web searchers, but very inexperienced digital video searchers, less experienced than the students and staff. The analysts' expertise was in mining text sources and text-based information retrieval rather than video search. They had no prior experience with the Informedia interface or data under study and no connection with CMU or the NIST TRECVID community.

4. INTERFACE CAPABILITIES FOR TRECVID AND EXPLORATORY SEARCH TASKS

Today's commercial video search engines often rely on filename and accompanying text sources to accomplish video retrieval functions[8]. Users issue text queries to retrieve nonlinguistic visual imagery. The image retrieval community has focused on content-based indexing using pixel-level image attributes like color, texture, and shape[8, 9], where users supply a visual example, but the underlying low-level attributes makes it difficult for the user to formulate queries. To bridge this semantic gap, the multimedia research community has invested in developing a Large-Scale Concept Ontology for Multimedia (LSCOM), whereby semantic concepts like "road" or "people" can be used for video retrieval[10].

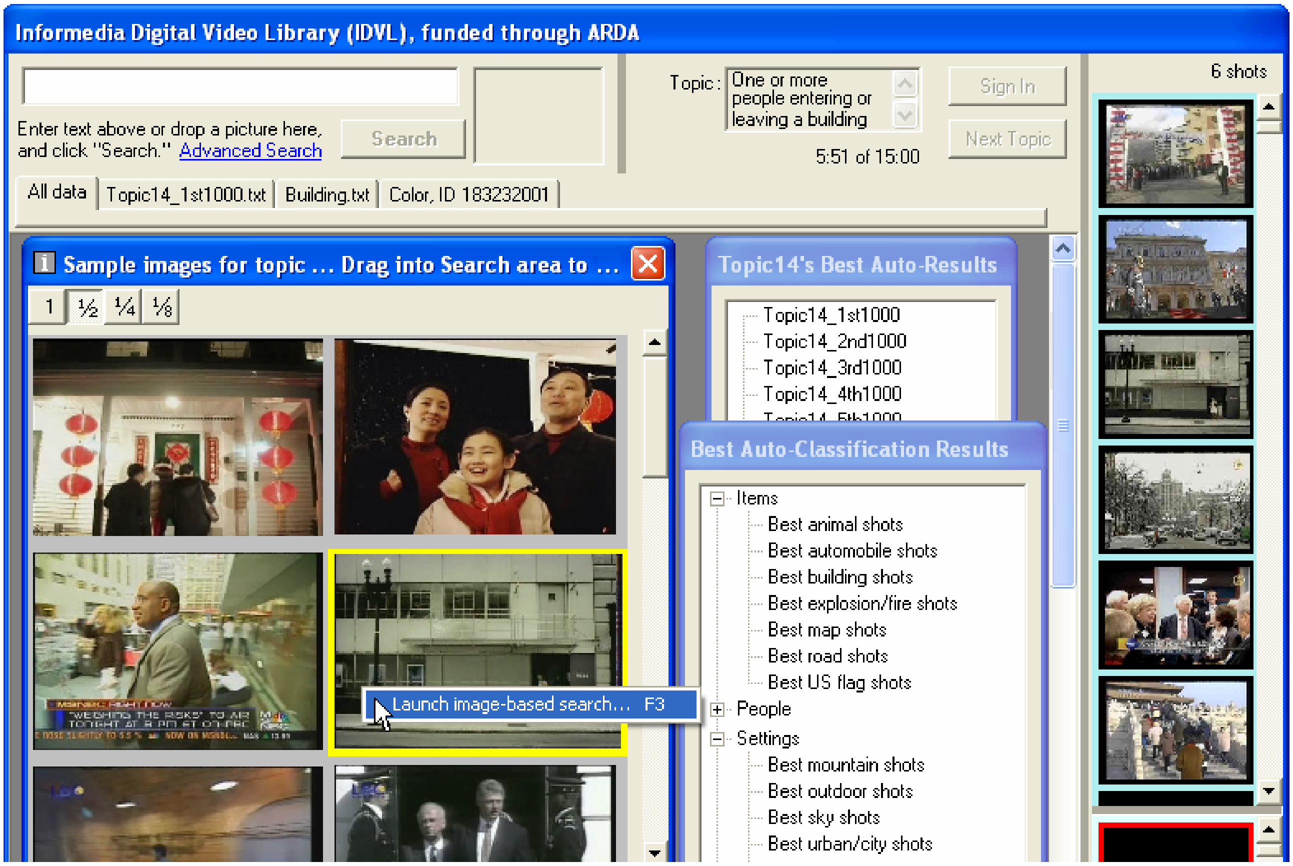
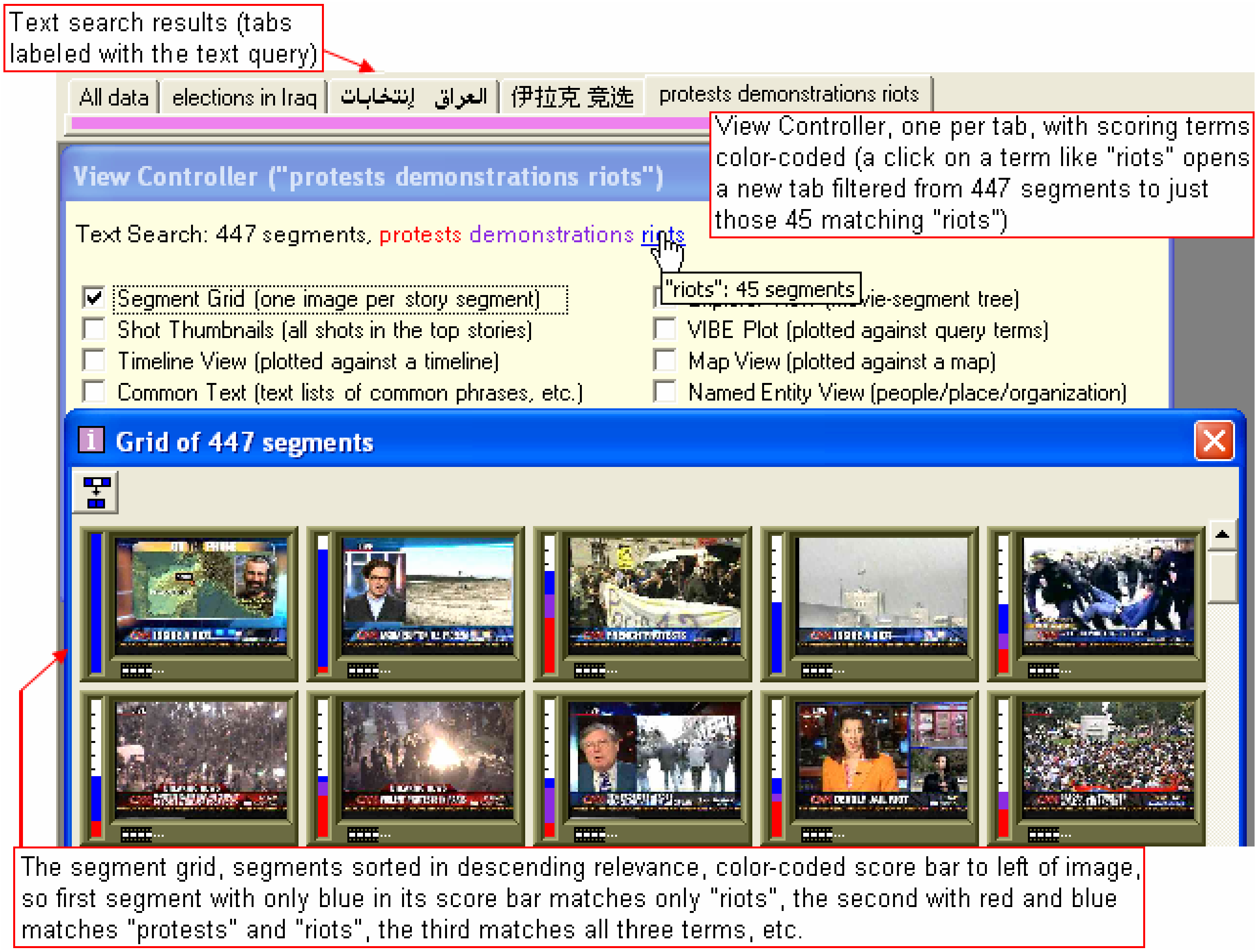


Figure 1. Interface used by analysts for TRECVID search topics (15 minutes/topic), with the goal being to mark relevant shots addressing the topic (which get collected and displayed in the pane at the extreme right).

These three access strategies, query-by-text, query-by-example, and query-by-concept, have been used by the Carnegie Mellon Informedia video search engine[1, 11] and the MediaMill video search engine[8] for a number of years, with these systems scoring best for all of the TRECVID interactive video search evaluations since the task inception in 2002[4]. The user studies reported here fold in a specialized form of query-by-concept for the TRECVID topics: the ranked shot output of the fully automated search, which we label "query-by-best-of-topic" since this is a topic-specific shot list. For example, for the topic "one or more people entering or leaving a building" a fully automated process produced a ranked list of 5000 shots, with the ENVIE interface partitioning this list into the first 1000, second 1000, …, fifth 1000, and then within each set of 1000 reordering the shots to preserve shot ordering within story segments. All shots from the same video segment as the first shot in the set are promoted to the first slot within the storyboard, then the next shot from a different segment in the ranked list is processed in the same way to promote its segment shots, until all the shots in the set of 1000 are covered. In this way, the earlier claim that Informedia storyboards preserve temporal ordering within segments is respected.

A sample screen shot as seen by the analysts for the TRECVID tasks is shown in Figure 1. Addressing the listed topic, the analyst might issue a text search on “enter exit” or perhaps a specific building like “Capitol”, issue a color-based search using a thumbnail, browse the best “building” shots, or browse the best 1000 automatically returned for this topic “one or more people entering or leaving a building.” The analysts had one 15-minute timed practice run with which to get familiar with the system, and an abbreviated paper User’s Guide to refer to as well, just as the CMU students and staff were given in the TRECVID 2005 study conducted with the ENVIE system[11]. The full exploratory interface, without the need for simplification brought on by the 15-minute TRECVID topic time limit, is illustrated in part in Figures 2 and 3.


Text search results (tabs labeled with the text query)

View Controller, one per tab, with scoring terms color-coded (a click on a term like "riots" opens a new tab filtered from 447 segments to just those 45 matching "riots")

View Controller ("protests demonstrations riots")

Text Search: 447 segments, protests demonstrations riots

"riots": 45 segments

☑ Segment Grid (one image per story segment)
☐ Shot Thumbnails (all shots in the top stories)
☐ Timeline View (plotted against a timeline)
☐ Common Text (text lists of common phrases, etc.)
☐ Explorer View (movie-segment tree)
☐ VIBE Plot (plotted against query terms)
☐ Map View (plotted against a map)
☐ Named Entity View (people/place/organization)

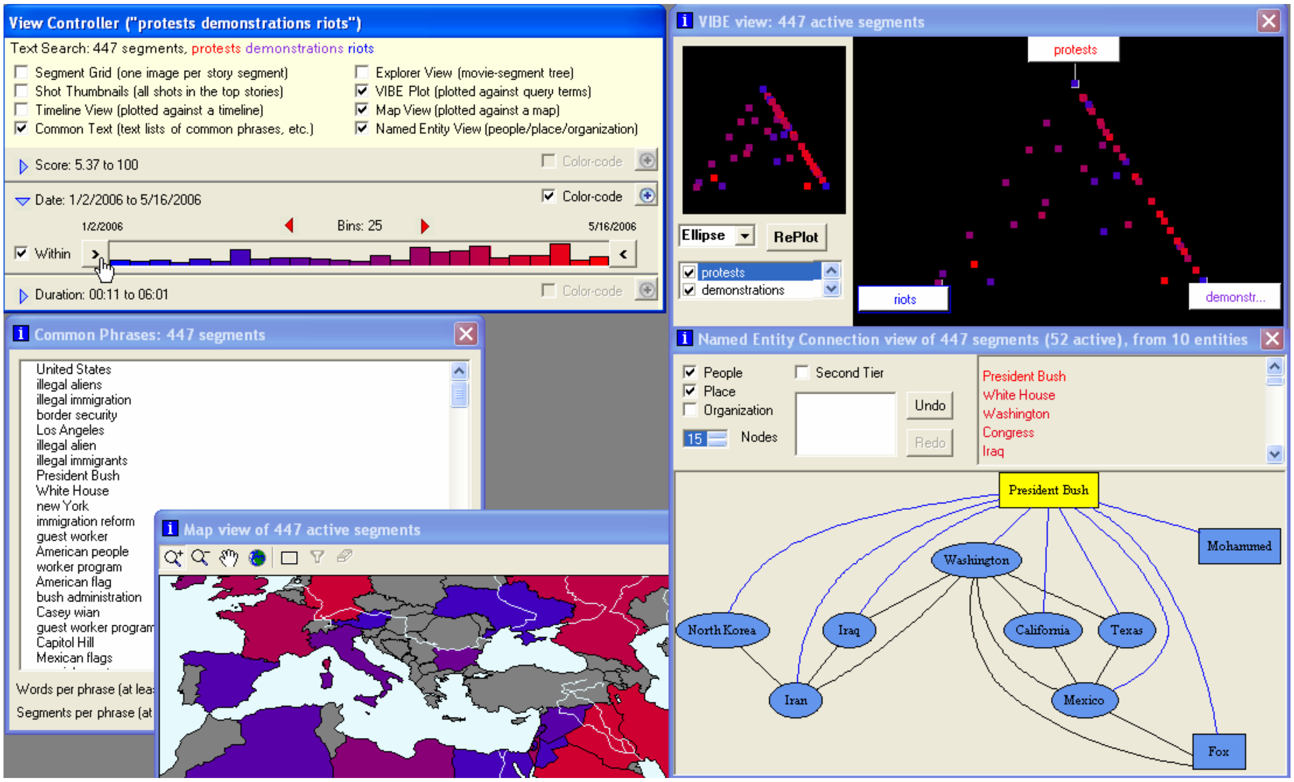
Grid of 447 segments

The segment grid, segments sorted in descending relevance, color-coded score bar to left of image, so first segment with only blue in its score bar matches only "riots", the second with red and blue matches "protests" and "riots", the third matches all three terms, etc.

Figure 2. Example of View Controller within ENVIE video set display.

For the TRECVID 2005 and 2006 tasks, a “Shot Thumbnails” view (i.e., storyboard) is used exclusively to present results back to the user, as this view best addresses the need for shot-based retrieval based on visual characteristics embodied in the TRECVID search tasks. The ENVIE system was designed to provide much greater functionality than just shot-based retrieval support, however, with a great deal of foundation Informedia interface work going into the design and development of video browsing and summarization "views" of sets of video data[1, 11, 12]. The ENVIE system represents the metadata for a video set produced by a query action (e.g., query-by-text, query-by-example, query-by-concept, query-by-best-of-topic within a TRECVID context or query-by-geographic-map outside of TRECVID context) in XML, and presents the video set in a tab along with a "View Controller" manipulated by the user to determine how that XML is displayed in the interface. The View Controller allows the video set to be rendered in different ways, i.e., checking a view "on" displays an additional window into the video set held in the tab. These windows each specialize in highlighting particular attributes of the video set, and many let you filter down to a subset in specialized ways. The advantage of the different views is to let you explore the set of video data in varied means, rather than restrict you to a thumbnail grid as shown in the storyboard of the Shot Thumbnails view. Figure 2 shows an ENVIE screen shot with tabs, a view controller for the tab produced by a text query "protests demonstrations riots", and the Segment Grid view emphasizing video story segments rather than shots, along with additional overlays of score and term contributions.

The Timeline View emphasizes time in a scatter plot view with time as the x-axis, and the y-axis by default the relevance score but also supporting other attributes like "video duration." Each plotted box represents at least one but possibly many video segments. Similarly, the "VIBE Plot" is a VIsualization By Example plot in which query terms (words for text queries, matching cities/states/countries for geographic searches) act as anchors for the plot, and then distributes the video segments on the plot based on their relative score from each of the anchors. If you right-click on a plot-box that represents a single segment, you can play its video, show its storyboard, or show its movie info, just like you can do with thumbnail representations. The Common Text View displays the most frequent phrases occurring in the text metadata associated with a video set, the Explorer View displays a hierarchy of broadcasters-shows-segments, the Map View shows geographic distributions, and the Named Entity View showcases temporal entity relationships. Typically, a user would manipulate at most a few of these at once to best use limited screen real estate, but for the sake of condensing the figure space used in this paper, 4 views are shown simultaneously in Figure 3.


View Controller ("protests demonstrations riots")
VIBE view: 447 active segments
Common Phrases: 447 segments
Map view of 447 active segments
Named Entity Connection view of 447 segments (52 active), from 10 entities
protests
riots
demonstr...
Ellipse
RePlot
Undo
Redo
President Bush
Mohammed
Washington
North Korea
Iraq
California
Texas
Iran
Mexico
Fox

Text Search: 447 segments, protests demonstrations riots

☐ Segment Grid (one image per story segment)
☐ Shot Thumbnails (all shots in the top stories)
☐ Timeline View (plotted against a timeline)
☑ Common Text (text lists of common phrases, etc.)
☐ Explorer View (movie-segment tree)
☑ VIBE Plot (plotted against query terms)
☑ Map View (plotted against a map)
☑ Named Entity View (people/place/organization)

Score: 5.37 to 100 ☐ Color-code
Date: 1/2/2006 to 5/16/2006 ☑ Color-code
1/2/2006 Bins: 25 5/16/2006
☑ Within
Duration: 00:11 to 06:01 ☐ Color-code

☑ protests
☑ demonstrations

United States
illegal aliens
illegal immigration
border security
Los Angeles
illegal alien
illegal immigrants
President Bush
White House
new York
immigration reform
guest worker
American people
worker program
American flag
bush administration
Casey wian
guest worker program
Capitol Hill
Mexican flags

Words per phrase (at lea
Segments per phrase (at

☑ People
☑ Place
☐ Organization
☐ Second Tier
15 Nodes

President Bush
White House
Washington
Congress
Iraq

Figure 3. Example of multiple views into the video set, with each view supporting unique opportunities for filtering, e.g., to just videos mentioning certain entities, phrases, map regions, or search terms.

Note that all views can also be controlled through the use of dynamic query sliders[13]. In this figure, if the mouse (hand cursor over the Date Slider) drags the date slider left boundary now at January 2, 2006 over to March 1, 2006, then all representations for metadata that only occur for January and February segments in the video set would drop out of the shown Common Text, Map, VIBE, and Named Entity views. That is, query slider manipulations immediately and directly affect the views being shown. Similarly, the information visualization technique of brushing allows the interrelationships between the emphases in the different views to be explored. In Figure 3, if the common phrase "border security" is highlighted in yellow by mousing over the phrase, then for example all the plot points in the VIBE view for segments containing "border security" will also be "brushed' with the color yellow. The rich dynamics supported by the video set views, information visualization techniques, and operations in support of the Visual Information Seeking Mantra "overview first, then zoom and filter, details-on-demand"[14] were communicated to the analysts participating in the ENVIE assessment in 3 ways: a demonstration session, group interviews with the ENVIE development team, and through a 35-page paper ENVIE User's Guide with annotated screen shots like that of Figure 2.

5. TRECVID 2005 RESULTS

The within-subjects study conducted with TRECVID 2005 topics had two goals: (1) confirm that situation analysts, like the CMU students working against the same topics with the same interface[11], made use of all provided query strategies with a resulting good performance on tasks; (2) through a within-subjects experiment, quantify and qualify the differences between simplified multimedia retrieval systems where only keyword text search is provided, versus the full Informedia system offering query-by-text, query-by-example, and query-by-concept. The within-subjects design has the advantages of holding subject variables constant (e.g., an outstanding video searcher contributes across all treatments), increasing statistical power by reducing random variation. The disadvantages of within-subject designs include the lasting effects of treatments and other time-sensitive effects like fatigue. We control for time-ordered effects by counterbalancing the systems under study so that half of the time subjects see one system variant first and half of the time it is the other system first.

We created two systems with nearly identical user interfaces but with one system, Text-Only, being a "text-only" system making use only of the speech narrative for query-by-text. In the Full system, query-by-text as well as query-by-example image color similarity search, query-by-concept search using the 39 LSCOM-lite concepts[10], and topic-dependent query-by-best-of-topic search, each shown in Figure 1, were available. The topics and systems were counter-balanced so that in a first session with 4 topics, the first 2 topics were given as Text-Only or Full and the second 2 topics in the other system, with the analysts each working through a second session of 4 topics in which the system order was reversed.

The analysts scored well on the TRECVID 2005 topics, especially since the six analysts reported no prior experience at all with video search systems. Their mean average precision (MAP) of 0.251 when using the Full system correlates well with the 4 student runs' MAP in a TRECVID 2005 study[11] of 0.253 through 0.286 with the same system. These student runs produced the highest MAP for TRECVID 2005 interactive search conducted by users outside of the system development teams[4, 11]. Looking at the average precision across the 24 topics shown in Figure 4, the analysts underperformed compared to the students on three "easy" tasks where the students performed well: topics 156 ("tennis players"), topic 165 ("basketball players") and topic 171 ("soccer goal"), the three sports topics. In later discussions, the analysts indicated disdain and perceived irrelevance for these sports-centered topics as they did not correlate well with their work, so it is not surprising to find that the analysts perhaps did not take answering these topics as seriously as the others. If the three sports-related topics are ignored, the MAP for the four student runs are 0.249, 0.228, 0.242, and 0.201, with the analyst run having a MAP of 0.248.

The MAP across all 24 topics for the analyst runs with the Text-Only treatment was 0.204 while the MAP for their Full runs was 0.251. The MAP for the 21 non-sports topics for Text-Only was 0.178 while the MAP for Full was 0.248. The context of the within-subjects study within a series of experiments exploring human-centered computing for video retrieval is discussed in detail in a separate paper currently under review, with the conclusion that even analysts having very high text-search experience and no video-search experience make use of, prefer, and perform significantly better with the Full system rather than the Text-Only system. The questionnaire responses support the conclusion that the full-featured ENVIE system was strongly preferred over a simple text-only video retrieval system. When using the Full system, the analysts took advantage of other access mechanisms beyond text, supplying shots in their answer set 20% of the time from query-by-text, 21% query-by-concept, 23% query-by-example, and 36% query-by-best-of-topic.

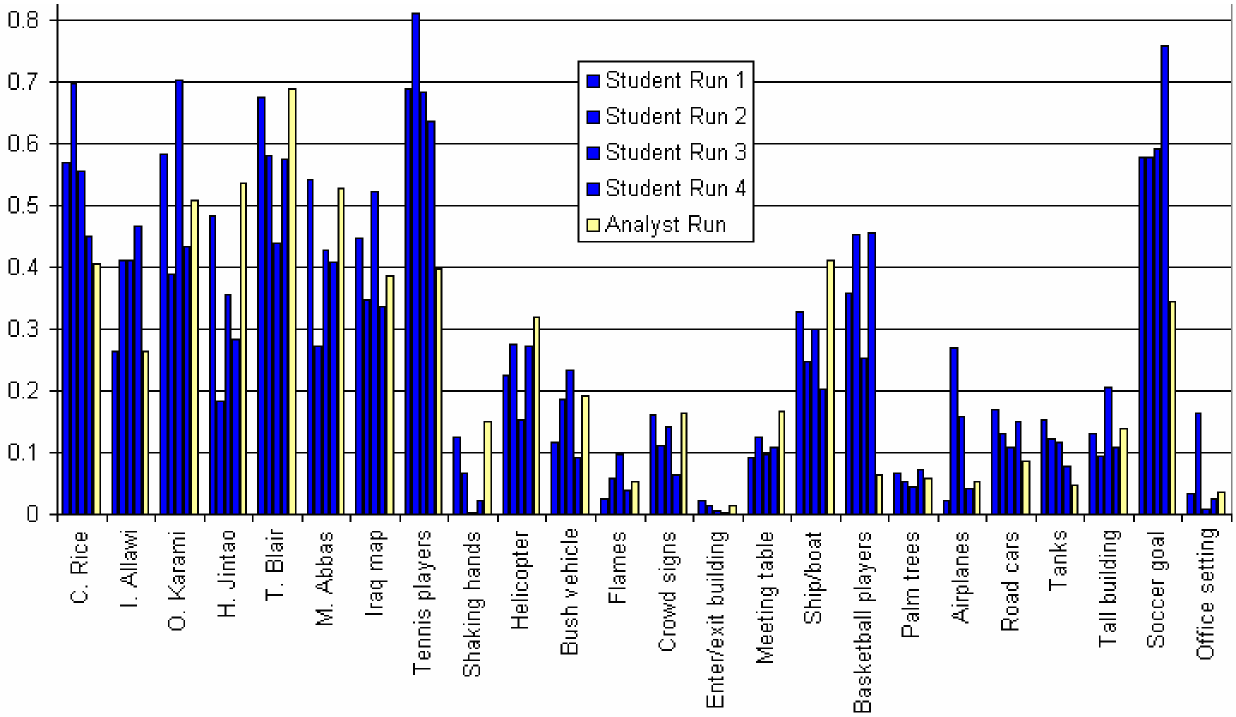


Figure 4. Average precision across 24 NIST TRECVID 2005 topics for 5 different ENVIE runs.

6. TRECVID 2006 RESULTS

The analysts began their two-day ENVIE assessment with TRECVID 2005 timed topics, and concluded with TRECVID 2006 timed topics the next day. In retrospect, too much was asked of the analysts in a compressed amount of time, and fatigue was clearly a factor for their TRECVID 2006 runs. They did not interact to the same level as did developers of the ENVIE system under the same 15 minutes/TRECVID topic time constraints, as expected from prior TRECVID reports[4], but what was not expected was the level of drop-off in activity. There also was a mistake attributable to fatigue: one analyst answered topic 174 (tall buildings) instead of 194, so we have no analyst data for topic 194, Condoleezza Rice. Figures 5 and 6 show the TRECVID shot count for each of the 24 TRECVID topics, with the online TRECVID materials[4] describing the meaning of the topics further beyond these terse text labels on the x-axis. It is interesting to note that again, as was done with TRECVID 2005 sports topics, the analysts disregard the sports topic (Topic 195 about soccer goal post) and contribute very few shots.

For TRECVID 2005, the comparison users were CMU students who also had never seen the ENVIE system before, but in Figures 5 and 6 the comparisons are being made to Informedia researchers, i.e., ENVIE system experts, who have experience in what concepts are likely to have reasonable accuracy (e.g., "outdoors", "roads", "buildings") and are highly motivated to perform at fast interaction speed to showcase ENVIE functionality, as they are the ENVIE developers. One CMU expert had the same ENVIE interface as the analysts; the other two used a restricted "query-by-best-of-topic only" interface. The analysts do not have the same motivation or experience as these experts, so it is expected that they will not perform to the same levels as the system developers, just as students did not perform to the same levels as developers in Informedia user studies for TRECVID 2004 and 2005 interactive search tasks[1, 11, 15, 16]. All four runs had available the ability to mark shots into two pools: those likely correct and relevant put into a "yes" pool in the shot collector pane shown to the extreme right of Figure 1, and those possibly relevant put into a "maybe" pool. Figure 5 shows just the counts of "yes" shots marked by the users, with Figure 6 showing "yes" plus "maybe" counts.

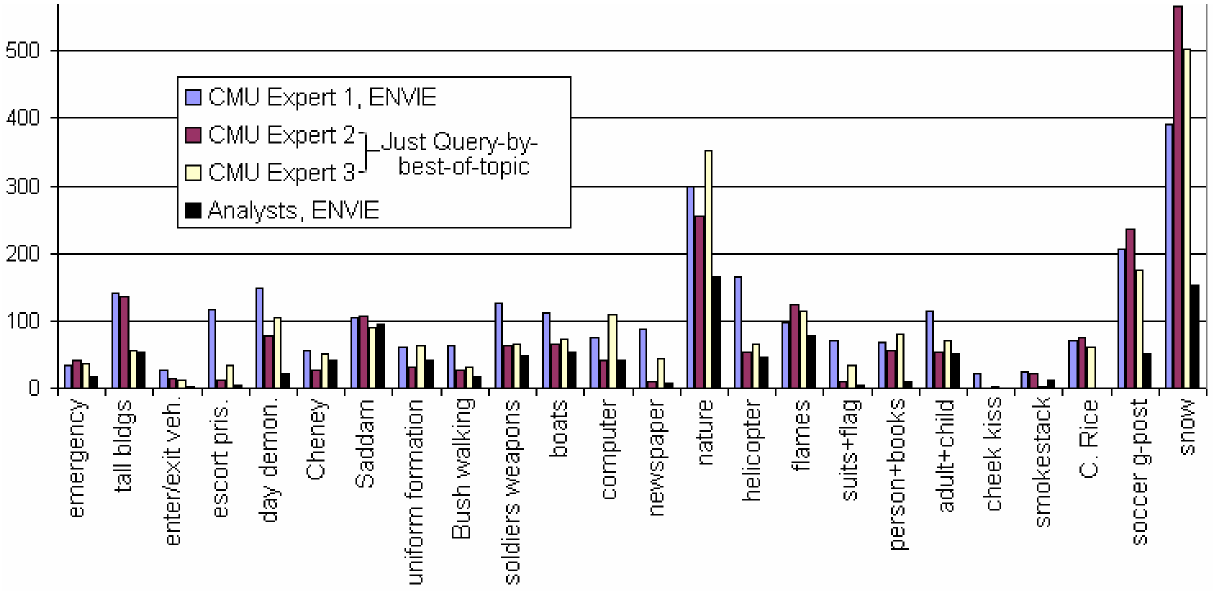


Figure 5. Number of TRECVID "Yes" shots per TRECVID 2006 topic across 4 runs.

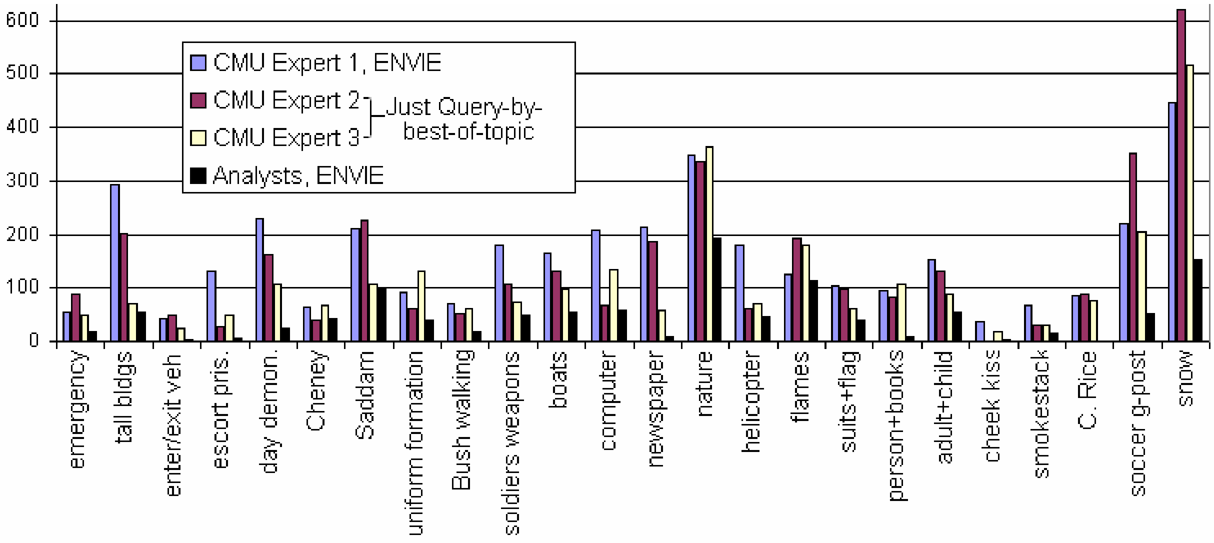


Figure 6. Number of TRECVID "Yes" and "Maybe" shots collected per TRECVID 2006 topic.

In fact, the questionnaires show the difference in attitude toward the TRECVID topics: the CMU experts want to maximize recall knowing that their precision is quite good and so are always eager to find even more relevant shots for any particular topic. The analysts are content in finding a number of relevant shots (good precision), and do not feel the urgency to find even more ones. Three questions asked after every topic was addressed were:

- I found that it was easy to find shots that are relevant for this topic.
- For this topic I had enough time to find enough answer shots.
- For this particular topic I was satisfied with the results of my search.

With a 5-point scale of (1=”Not at all”, 5=”Very much”) the analysts’ responses for these three questions averaged to 3.83 (fairly easy to find shots), 4.21 (had more than enough time), and 4 (satisfied with results). The CMU expert’s averages were 4.17 (easy to find shots), 2.46 (not enough time), and 2.75 (not satisfied with results) for the same full-query-access ENVIE system, because the CMU expert wanted to get hundreds rather than tens, and even when hundreds were collected, still wanted more time to find additional shots when the topics were generic enough that the perception was that there were more to be found. For example, the analyst was highly satisfied with the collected set and felt there "very much" was enough time after collecting 18 shots for emergency vehicles (the first topic), or 152 shots for snow, while the CMU expert was frustrated by the 15-minute time running out and was barely satisfied (rating them 2) for collecting 34 emergency vehicle shots, or 392 snow shots. Hence, these user groups considered their answer sets differently. Only one of the six analysts ever rated a topic at 1 "Not at all" or 2 on the 5-point scale for either having enough time or being satisfied with the search results.

The TRECVID organizers should take note of this phenomenon and address it in subsequent annual forums if ecological validity for situation analysis work is to be a high priority: a task to find up to 1000 relevant shots may not be well-grounded, with analysts content to stop once tens of relevant shots have been identified. This lack of concern by the analysts to identify hundreds of shots per topic resulted in lower average precision across the TRECVID 2006 topics: the performances graded by NIST pooled truth are shown in Figure 7. The mean average precision for CMU Expert 1 with exactly the same system was 0.303, the 2 CMU experts using query-by-best-of-topic only systems scored MAPs of 0.184 and 0.250, and the analysts' run had a mean average precision of 0.150. If, instead of giving the analyst run a score of zero for the skipped "C. Rice" topic, we compute MAP across the remaining 23 topics, the values are 0.289 for Expert 1, 0.165 and 0.229 for CMU Experts 2 and 3 with the restricted system, and 0.157 for the analysts.

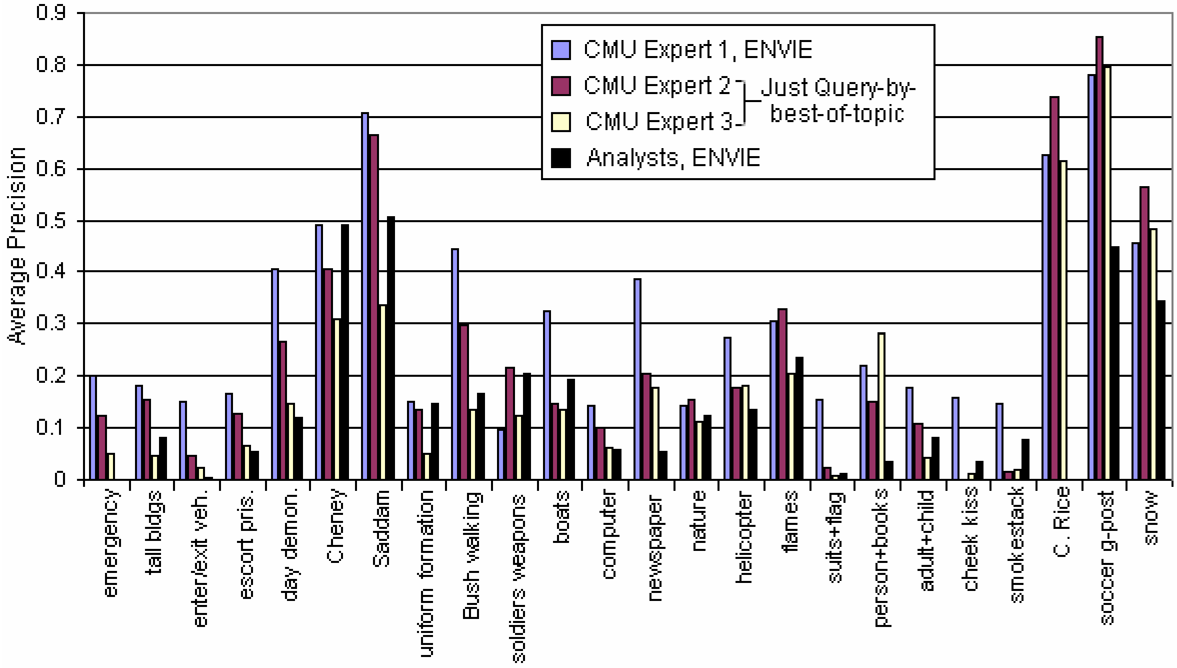


Figure 7. Average precision for TRECVID 2006 topics across 4 different runs.

Figure 8 shows that the ENVIE system allows for thousands of shots to be reviewed, by both highly motivated and experienced ENVIE researchers as well as by the analysts new to the system and without the same pride of ownership in the system, within the 15 minute time period allowed for addressing a TRECVID topic. The interface has been designed for efficient access and review of huge amounts of imagery as surrogate views into great volumes of underlying video data, and these TRECVID metrics, in conjunction with the interview data, confirm that the analysts recognized and appreciated ENVIE's capabilities in this regard.

Within the ENVIE system, for TRECVID 2006 there existed enhanced query-by-concept functionality based on machine learning and algorithmic improvements for automated visual classification[17]. For each topic, e.g., "helicopter" or "snow", the system would automatically compute the best-shots list for that topic, lists with these query-by-best-of-topic sets accessible by the ENVIE users. Both the CMU expert with ENVIE and the analyst user pool made heavy use of these best-of-topic sets, as shown in a breakdown of interactions in Figure 9. The transaction logs show that in spite of the six analysts' impressive experience with text-based analysis and retrieval, and self-reported lack of any experience with video retrieval systems, they did make use of the ENVIE information access strategies besides just query-by-text. The expert used query-by-text 16% of the time, the analysts 20% of the time. (By contrast, CMU Experts 2 and 3 reported in the earlier figures had access only to best-of-topic, so by design their interactions were forced to be 100% "query-by-best-of-topic" as part of a separate study isolating the benefits of that access strategy.)

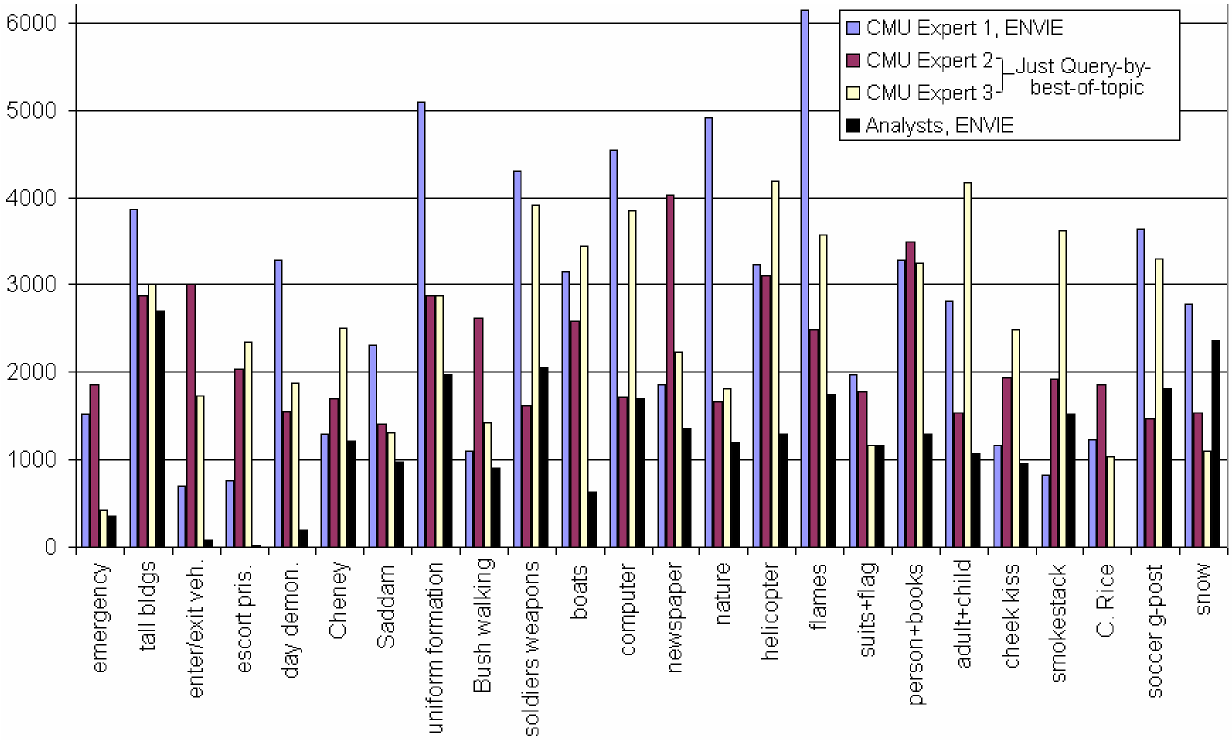


Figure 8. Number of checked shots (captured, or skipped during capture), across TRECVID 2006 topics.

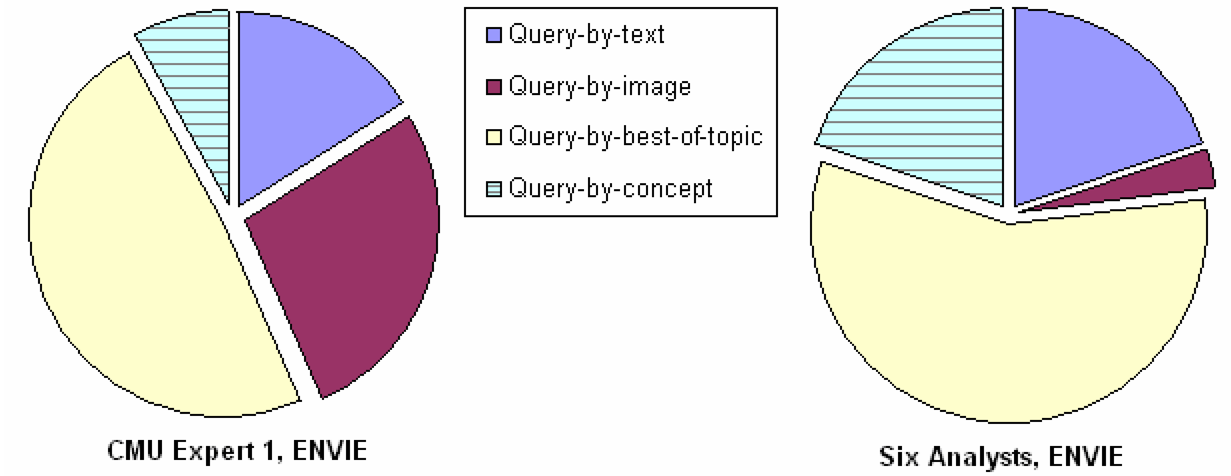


Figure 9. Percentages for sources of shots captured into "yes" set for the TRECVID 2006 topics, indicating use of ENVIE access mechanisms beyond text search (query-by-text a minor access strategy).

7. EXPLORATORY SEARCH AND INTERVIEWS

The TRECVID 2005 and 2006 sessions provided quantitative and qualitative metrics supporting the ENVIE design as productive for shot-based retrieval tasks given an expressed information need, the TRECVID topic. Analyst activity is anticipated to be more creative and exploratory as well, where the information need is discovered and evolves over time based on interplay with data sources. Evaluating tools for exploratory, creative work is difficult, with multi-dimensional assessment strategies recommended for assessment[18], and hence we employed transaction logs, group and individual interviews, and think-aloud sessions. The procedural recommendation is to encourage such multi-dimensional assessment in the future, as the techniques complement one another: transaction logs capture what was done but not why, with think-aloud protocol providing the insights into analyst reactions and strategies. Group sessions can save time, but can be dominated by a minority of participants because of rank, job status, or personality, with individual interviews and think-aloud sessions offering every user the private opportunity to comment on system features and use. For the ENVIE assessments reported here, the individual feedback sessions were more informative, with fewer comments made during group sessions (e.g., three analysts stayed completely quiet in group sessions). In general, ENVIE was found to be capable of presenting vast amounts of video and imagery, search such data very efficiently through multiple means, and allow easy capture of subsets of information. Overall, the six analysts did a great deal of work with the ENVIE system in an extremely short testing period. During their interactions with ENVIE they collectively logged 1433 video plays and 433,031 shot scans, in addition to hours spent with questionnaires, interviews, and other interface widgets and presentation schemes. From the interview data, a number of positive remarks were made, including the following:

> "This is the fourth information retrieval system I have evaluated …and it is the easiest to learn and use, provides great functionality…. I could familiarize myself with the interface features quickly and use them to accomplish both the TRECVID tasks and exploratory tasks. The 'best-of' sets, timeline, map search, all were useful."
>
> "Fast system response time, great speed in searches. Quick presentation of great volume in imagery. Fast interaction to get to synchronized video point corresponding to thumbnail or query. Ability to learn how to use system with only a few sessions' experience."
>
> "Great speed. Lots of imagery/video presented in little time. Images clear and informative, with fast synchronized video access. Other views (named entity, map, etc.) useful for exploratory topics…."

Exploratory topics, especially when supplied by the analysts themselves since the instructions allowed for a "fill in your own topic search", were often difficult to satisfy. Four reasons are offered for the difficulty in finding relevant video for some exploratory topics with ENVIE:

1. The provided test corpus was small (240 hours of news covering January-May 2006), so some topics, e.g., an analyst-supplied "natural environments affected by global warming", did not have much support in the corpus. For this topic for example, the same 3 "stock CNN topics" on bears, glaciers, and hurricanes dominated the found support materials. If the corpus covered more time or sources, one could envision materials on volcanoes, tsunamis, El Niño, etc., now being available to draw into support materials.

2. The training time given the analysts for using ENVIE was very brief, an hour demonstration and distribution of a 35-page ENVIE User's Guide which many analysts never found the time to read (at least two times, analysts self-reported on questionnaires that "oops, that feature is there, just see it now in User's Guide"; in interviews, the most common remark was that there was too little time to get into the advanced features of the ENVIE system and learn how to use features like shot filtering based on concepts and visual filtering, and that they wished they had more time to investigate such features). As a result, transaction logs show that the most frequent operations performed were text search (the analysts had a great deal of expertise in text retrieval systems and mining text information repositories), storyboard/thumbnail browsing, and video playing, with other views like the named entities connection graph, map visualizer, timeline, and the visualization-by-example scatter plot (see Figure 3) touched upon but not used in detail.

3. The tasks were unstructured (on purpose, to keep them "exploratory" rather than "find X" specific topic search), and also forced into a two-day agenda. Some topics might require longitudinal analysis with an ENVIE/reflection/back-to-ENVIE iterative loop; other topics might not be answerable in an hour's interaction, and other topics might not be well-grounded in an individual analyst's expertise or expected situation analysis activity.

4. The exploratory ENVIE interface discussed in the 35-page ENVIE User's Guide was feature-rich, but also unfortunately hid the query-by-concept feature that was exposed better in the streamlined interface used for TRECVID 2005 and TRECVID 2006 tasks. Likewise, this first use of the ENVIE system by the analysts produced a number of comments on the numerous but relatively simple-to-address interface features (e.g., windows management issues of cropping, layout, size) that were lacking or incomplete because ENVIE had been developed as an operational prototype rather than a finished product.

The following actions could be taken to address these points and make future assessments of exploratory video analysis interfaces more valuable (addressing the same points 1-4 enumerated above, respectively):

1. To better assess the utility of an exploratory system allowing for analysts to supply their own queries, a broader, larger, more comprehensive test corpus would be ideal. For example, the Informedia research group could test against a CNN corpus of the past seven years. Ideally, corpora of foreign news covering multi-year time spans could be folded in.

2. The same analysts could use the tested system over a time period of weeks or months, allowing for the system to be learned and specialty features within the tool appreciated. Shneiderman and Plaisant discuss additional benefits for longitudinal studies[18], e.g., the opportunities to make use of a specialty tool such as perhaps ENVIE's named entity viewer crossed with outdoor-road-people shot filtering may occur rarely but when it does the specialty tool shows incredible merit, so a longer transaction period is necessary to ever capture such a rare event.

3. Better grounded exploratory tasks from actual intelligence activities conducted within the analyst's work facilities would provide the ideal case studies for evaluation, but might introduce a number of logistics issues.

4. With a single, first iteration through an operational prototype, the most glaring problems are easily, repeatedly identified and those problems can be fixed for the benefit of the analysts and improvement of the system on its evolution toward a deliverable system. A much higher percentage of analyst time was spent fighting through and reporting on interface concerns that could be fixed to produce more valuable user feedback in follow-up iterations (where both the ENVIE system would be improved based on analyst feedback, and the analyst would have greater system experience to draw from in dealing with the video analysis system). An assessment as reported here should actually be only the first step of what should ideally be an *iterative* assessment-development process.

Despite the need for more exhaustive testing of ENVIE as an exploratory search tool, the analysts' feedback shows that much video data can be accessed and surveyed under tight time constraints by people trained in text-based retrieval systems, since having powerful text-based access into video corpus is an expected first step into the corpus by analysts with such text-based experience. Further, the initial "setting" of the ENVIE system is quite important, especially for users new to the tool, as those initial settings may not be strayed from much at the start and ideally could save the user a great deal of work if tuned to the user's profile (e.g., for a traffic situation analyst, automatically suppressing all television studio shots and emphasizing all shots of roads and vehicles in the field).

The think-aloud protocol[19] asked the analysts to verbally comment on their actions, their reactions, what confused them and what differed from the expected, as they worked through these topics with ENVIE during a twenty minute period: "find street shots of Baghdad"; "find Chinese news sources showing street shots of Baghdad", "find vehicles in street shots of Baghdad", "find people and vehicles in street shots of Baghdad." These aural records provide understanding into trends seen with exploratory topic transaction logs. The analysts start with tools they are comfortable with: text search primarily, including accessing ENVIE's "advanced text search" capabilities significantly more than CMU students ever did in prior testing. If too many results are returned, the analysts in general are willing to reissue more focused refined text search rather than use a visual-based filtering tool on top of already collected results, hence returning to their text search expertise. Rather than generalize all the actions together, though, the protocol shows that each analyst has individualized preferences and biases. Video retrieval tools should be flexible enough to complement and extend analyst's skill set: some trust results too much, others too little. Some use visual filters and take advantage of other ENVIE advanced features even with less than a day's exposure to the system, others not. Some do image search with overly optimistic expectations, others with realistic expectations, and others not at all. All assume the existence of a state-of-the-art text search interface, so when "simple" things like Baghdad spelling correction for "Bagdad or "Bahgdad" is not provided, they were confused or annoyed.

8. CONCLUSIONS

User studies conducted with TRECVID topics and data have a vast head start over studies conducted by individual research groups against privately grown, often much smaller corpora, 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. A 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 documents, the differences in expert and novice search behavior, the utility of transcript text for news topics, and the employment of concept filters (e.g., include or exclude all shots having the face feature or "outdoors" feature) to reduce the shot space. This session with the analysts offered the opportunity for improved ecological validity with respect to intelligence analysis: rather than use university students and staff as surrogate analyst users, the analysts themselves took part in the TRECVID interactive search task experiments. The TRECVID studies confirmed the utility of multiple access strategies to the news video beyond just query-by-text, in agreement with usage patterns witnessed in CMU students' and Informedia developers' runs against the TRECVID topics. Query-by-example, query-by-concept, and query-by-best-of-topic collectively were used much more than query-by-text, despite the analysts' high level of expertise with text retrieval and inexperience with video retrieval. The TRECVID experiments here also show that for improved ecological validity with a situation analyst population, sports topics should be dropped as this audience considers them unrepresentative of the work they do. Importantly, the analysts also consider collecting tens of relevant shots a successful run, rather than hundreds. Further focus groups could be consulted to determine whether indeed the use of average precision across 1000 shots is not representative of a real-world interactive search task, with average precision across 100 or 200 being more realistic.

Of course "video search" is much broader than the shot-based retrieval from news corpora emphasized in recent TRECVID years. Video search activity is creative and exploratory as well, where the information need is discovered and evolves over time based on interplay with data sources. Evaluating tools for exploratory, creative work is difficult, as acknowledged by Shneiderman and Plaisant[18]. TRECVID may very well broaden its scope to cover other issues in video search, for example the creative, exploratory discovery of materials in video corpora rather than seeking relevant materials for a known, expressed need. The assessment strategies should broaden as well, embracing the use of "Multi-dimensional In-depth Long-term Case-studies (MILC)"[18]. Ideally, MILC research could be conducted with representatives of a user community over time, to see changing patterns of use and utility as the people gain familiarity and experience with the system. In the term "Multi-dimensional In-depth Long-term Case studies" the multi-dimensional aspect refers to using observations, interviews, surveys, as well as automated logging to assess user performance and interface efficacy and utility. The in-depth aspect is the intense engagement of the researchers with the expert users to the point of becoming a partner or assistant. Long-term refers to longitudinal studies that begin with training in use of a specific tool through proficient usage that leads to strategy changes for the expert users. Case studies refer to the detailed reporting about a small number of individuals working on their own problems, in their normal environment. Longitudinal studies have been carried out in HCI and in some information visualization projects, but MILC proposes to refine the methods and expand their scope. The controversial question is how far video search system researchers can go in measuring the utility of their tools by the success achieved by the users they are studying, i.e., a way to keep technical developments in synergy with human needs.

This paper reports on a suite of HCI techniques employed to make the most of the analysts' time over two days with the ENVIE system, addressing the "multi-dimensional" aspect but falling short on other points. Without repeated investigations over a longer period of time, the users will be fatigued by pushing too much work into too compressed a time period (which contributed to lower TRECVID 2006 performance as those tasks were at the end of the activities as listed in Table 1), and users will not gain enough experience with the system to employ it optimally in addressing exploratory tasks. Ideally, MILC research could be conducted with the analysts over time, to see changing patterns of use and utility as the analysts gain familiarity with the system. Even so, the multi-dimensional techniques used in this two-day ENVIE assessment were an excellent initial step exploring the use and utility of the ENVIE tool for news video analysis and exploitation.

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 significantly enhanced through the efforts of Robert Baron and Bryan Maher.

REFERENCES

1. Christel, M. Evaluation and User Studies with Respect to Video Summarization and Browsing. *Proceedings of SPIE Volume 6073, Multimedia Content Analysis, Management, and Retrieval 2006,* E.Y. Chang, A. Hanjalic, and N. Sebe, eds., DOI 10.1117/12.642841.
2. 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.
3. Foraker Design. "Usability in Website and Software Design," http://www.usabilityfirst.com/. Accessed Nov. 2006.
4. National Institute of Standards and Technology. *Digital Video Retrieval at NIST: TREC Video Retrieval Evaluation (TRECVID)*, http://www-nlpir.nist.gov/projects/trecvid/. Accessed Nov. 2006.
5. 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.
6. Shatford, S. Analyzing the Subject of a Picture: A Theoretical Approach. *Cataloguing and Classification Q., 6*, 3 (1986), 39-62.
7. Kraaij, W., Smeaton, A.F., Over, P., and Arlandis, J. *TRECVID 2004 Proceedings*, http://www-nlpir.nist.gov/projects/trecvid/.
8. Snoek, C., Worring, M., Koelma, D., and Smeulders, A. Learned Lexicon-driven Interactive Video Retrieval. *Proc. CIVR 2006* (Tempe, AZ, July 2006), *Lecture Notes in Computer Science 4071*, 11-20.
9. Smeulders, A., Worring, M., Santini, S., Gupta, A., and Jain, R. Content Based Image Retrieval at the End of the Early Years. *IEEE Trans. Pattern Analysis and Machine Intelligence* 22(12) (2000), 1349-1380.
10. Naphade, M., Smith, J.R., Tesic, J., Chang, S.-F., Hsu, W., Kennedy, L., Hauptmann, A., and Curtis, J. Large-Scale Concept Ontology for Multimedia. *IEEE MultiMedia* 13(3) (2006), 86–91.
11. Christel, M., and Conescu, R. Mining Novice User Activity with TRECVID Interactive Retrieval Tasks. *Proc. CIVR 2006* (Tempe, AZ, July 2006), *Lecture Notes in Computer Science 4071*, 21-30.
12. 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.
13. 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.
14. Shneiderman, B. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. *Proc. IEEE Symposium on Visual Languages* (1996), 336-343.
15. 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.
16. 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.
17. Yan, R. *Probabilistic Models for Combining Diverse Knowledge Sources in Multimedia Retrieval*. Ph.D. thesis, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, 2006.
18. Shneiderman, B., and Plaisant, C. Strategies for Evaluating Information Visualization Tools: Multi-dimensional In-depth Long-term Case Studies. *Proc. BELIV'06 Workshop, Advanced Visual Interfaces Conference* (2006). Also available from the URL http://hcil.cs.umd.edu/trs/2006-12/2006-12.pdf.
19. 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.