Human Performance Measures for Video Retrieval

Gary Marchionini
University of North Carolina
100 Manning Hall
Chapel Hill, NC 27599
1.919.966.3611
march@ils.unc.edu

ABSTRACT
In this paper, we describe the challenges of assessing human performance during video retrieval episodes and describe several measures of human performance that have been used in developing visual surrogates for the Open Video Digital Library (http://www.open-video.org). These include two sets of cognitive performance measures that aim to assess human recognition and inference and a set of attitudinal measures that aim to assess user satisfaction with video surrogates.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: *Search process;* H.3.1. [**Content Analysis and Indexing**]: *Indexing methods;* H.5.1 [Multimedia Information Systems]: *Evaluation/methodology*

General Terms
Measurement, Performance, Experimentation, Human Factors.

Keywords
Video retrieval, exploratory search, information interaction, human-computer information retrieval.

1. INTRODUCTION
Information retrieval (IR) is always embedded in larger human endeavors such as work, learning, or entertainment. Ultimately, the success of an information retrieval episode is measured by how well it advances the primary endeavor, however, it is possible to make several simplifying assumptions and abstract the IR episode from the larger context for the purposes of system evaluation. The IR research and development community has adopted several simplifying assumptions over the past 40 years, and built effective assessment metrics that have propelled progress on system theory as well as system implementation. Key simplifying assumptions include focusing on document retrieval, adopting a binary relevance metric (assuming a retrieved item is either relevant or not), and removing human behavior from evaluation. Thus, IR has made dramatic progress by isolating retrieval from the human tasks and contexts and focusing on what algorithms and systems can do given one or more inputs (queries). There is increased interest in taking more of the context into account for retrieval, including the user and her task (see Spark-Jones for a recent, well-articulated statement of this trend, using the TREC tracks as examples. The prominence of IR in today's WWW-based environment supports a variety of IR sub-communities that relax the traditional simplifying assumptions. The question answering community aims to return answers rather than documents; the XML retrieval community aims to retrieve structured portions of objects (e.g., see [16] in this workshop); mainstream WWW search engine communities use mixed ranking models that combine different sources of evidence such as hyperlinks and term occurrence; media-specific communities aim to leverage the special properties of music, images, and video; and the information interaction community aims to consider the behaviors of people engaged in IR and indeed broaden the scope of IR to including exploration and learning contexts. Other communities are beginning to focus on specific types of retrieval tasks rather than on monolithic solutions that solve all retrieval tasks (e.g., see [4] in this workshop for examples of five different end-user oriented tasks in video retrieval.

This paper focuses on one set of approaches to include human factors in assessing specific aspects of video retrieval. We focus on human performance with video surrogates as a special aspect of video retrieval that influences overall retrieval. Video surrogates are kinds of metadata akin to 'glosses' or 'abstracts' that 'stand for' the full video object and are especially useful for the purposes of browsing and making sense of retrieved video objects[1]. Poster frames, storyboards, excerpts, and fast forwards are examples of surrogates we have evaluated. Thus, like classical IR evaluation, the measures are isolated from the full IR context in that they are specific to one aspect of retrieval and are applied in laboratory settings. They are unlike classical IR evaluation in that they focus on human performance rather than how well the retrieval system matches videos to queries. We first provide an overview of the challenges of video retrieval and exploratory search; then consider human performance and some ways that performance may be measured; next, we describe a set of human performance measures we have used in developing and assessing the Open Video Digital Library; and conclude with suggestions for new approaches to measuring human performance during video retrieval and use.

Conference'04, Month 1–2, 2004, City, State, Country.

[1] Metadata can be used for retrieval or to support understanding of retrieved objects. Visual surrogates may be useful for retrieval algorithms, however, our work has primarily focused on visual surrogates in result sets where people aim to make relevance judgments about whether to look further or download the video.

2. VIDEO RETRIEVAL EVALUATION CHALLENGES

Video content offers new opportunities and challenges for retrieval. Because video has multiple channels of representation, it is possible to leverage several media features. Visual features such as color, luminosity, texture, and shape are actively studied for the purposes of indexing. Likewise, audio features such as sound type (speech, music, natural sounds, silence), speaker alternation patterns, and decibel level are used to index video [7]. In spite of the enormous effort given to these different video features, the TRECVID evaluations have consistently shown the primary value of words for supporting video retrieval[2] [11, 13]. A key challenge for video retrieval systems is to integrate the different sources of evidence from these many features into indexing that helps people quickly find what they seek. Systems such as Fisclar and Intermedia leverage these multiple sources of evidence in different ways.

The video medium also offers two important challenges for video retrieval performance. First, the temporal nature of video and the volume of information contained per unit time invite segmentation that presumably improves retrieval through finer grained indexing as well as giving searchers advantages when managing and consuming video content. The same features that are used for indexing are also leveraged for the purposes of segmentation and it is inherently difficult to separate out the effects of segmentation from the effects of indexing when assessing video retrieval performance.

The second challenge is related to the representation of queries and the metadata that result from indexing. The work of the video retrieval community has begun to attack the latter representation by creating visual and audio metadata. We refer to these metadata as 'surrogates' and create four kinds of visual surrogates in the Open Video Digital Library (poster frames, storyboards, fast forwards, and short excerpts). The Informedia Project pioneered some of the most complex surrogates, called 'skims' more than a decade ago [15]. Our current work aims to add audio surrogates to the mix [1]. Thus, there is good progress toward creating non-textual representations for video.

Progress on query representation is much less evident. Current practice uses words as the basis for queries even when visual or audio metadata is available. The only practical alternative today, which is used by most video retrieval systems, including the major search engines, is query-by-example user interfaces that allow people to enter words and then ask for similar items, or begin with a diverse set of items and ask for more like one or a few of them. Presumably, the pervasiveness of mobile video-capable devices will stimulate audio queries and may eventually support using a real-time video stream as a query (e.g., find scenes like this). These challenges make video system retrieval performance difficult to evaluate as the techniques and user interfaces are in the early phases of practical development.

[2] This is mainly true regardless of whether the words come from transcripts, closed captioning, or automatic speech to text processes; however recent TRECVID results that used automatic machine translation from other languages yielded better results with visual features.

3. EXPLORATORY SEARCH EVALUATION CHALLENGES

In addition to the special opportunities and challenges that video content offer to retrieval evaluation, there is a trend toward broadening the kinds of retrieval that is evaluated. One of these trends aims to evaluate exploratory search rather than more precise known-item retrieval episodes. Marchionini [8] distinguishes information seeking episodes into three broad classes: lookup, learning, and investigation and argues that learning and investigation classes define exploratory search. Lookup includes fact retrieval, known-item retrieval, navigation (searching as an intermediate step to get to another source/website), transaction, verification, and question answering. Most of the classical evaluations in IR have focused on lookups because there is high probability of determining whether a retrieved item satisfies the query or not, thus allowing metrics such as precision and recall to be used. Additionally, the retrieval process itself can be measured using ratio-scaled metrics such as time.

Retrieval for the purposes of learning includes knowledge acquisition, comprehension, interpretation, comparison, aggregation, and socialization. Evaluating the results of queries is much more difficult in these cases and assessing the process is problematic as measures such as time are highly task dependent too cursory (e.g., learning may benefit from more time invested in the search process where conceptual connections and deeper understanding of the relationships among objects are desired).

Investigation goes beyond one or a few IR episodes and includes goals such as accretion of knowledge, analysis, exclusion/negation, synthesis, evaluation, discovery, forecasting/planning, and transformation. Evaluating IR activities devoted to these kinds of goals is difficult because progress may depend on failure (discovering dead ends) or may not be ascertainable until years later. Thus, to evaluate exploratory search, the simplifying abstractions and assumptions that are useful for classical IR cannot be made because the results are related to changes in the searcher rather than to the similarity of documents to queries. Evaluations of exploratory search must address the human factors (personal as well as global), the information content, and the retrieval system concurrently. A recent SIGIR workshop [18] on evaluation of exploratory search framed the problem and offered a range of qualitative and quantitative approaches to evaluating exploratory search. One of the key issues arising from efforts to include human behavior in retrieval is clearly specifying what is meant by performance—the system's actions or a human's actions. In our work, we are interested in the intersection of people using IR systems, specifically video retrieval systems.

4. HUMAN PERFORMANCE

It is useful to distinguish human physical, cognitive, and attitudinal characteristics as each contribute to human performance during search. Physical attributes such as reaction time, alertness, and endurance influence behavior and overall performance. Researchers observe both voluntary (e.g., large muscle movements, gross eye movements) and involuntary (e.g., heart rate, galvanic skin response, eye saccades, facial expressions) activity to assess performance. In the case of IR, physical activity is an adjunct to search performance, used to

make inferences about searcher preferences or to help explain search behavior. For example, one Open Video study used eye tracking to investigate how people use keyframes and text titles when viewing search results [5]. Such studies yield fine-grained evidence for very specific searching behavior but only provide partial insights into overall video retrieval performance.

Searching for information in an electronic environment is mainly a cognitive activity and thus assessment of cognitive variables are of primary interest except in the case of highly specialized or time-critical settings (e.g., target identification). One of the cognitive measures of human retrieval performance is search results accuracy---did the searcher find good information. In classical IR evaluation, the retrieved results are assessed by experts or panels of judges who assess the relevance of each result. It is relatively straightforward to ask searchers to say which videos retrieved are relevant to their needs, but these assessments tend to be specific to that person, task, and video corpus. Large-scale efforts to make judgments are expensive and the TRECVID community has aimed to apply the TREC text retrieval techniques of pooled relevance sets across sets of queries to address search accuracy. We have also experimented with measuring query quality [17] by having human judges count the number of pertinent concepts in a query episode (one or more queries expressed for a given topic). We found acceptable inter-rater reliability for the assessments and query quality measures may be a useful user-centered adjunct to results-oriented measures.

A more specific cognitive measure aims to assess mental load. Presumably, a system that requires less cognitive load is preferable to one that requires more load. There are many approaches to measuring cognitive load. Physiological measures such as galvanic skin response, heart rate, pupil dilation, and blood flow in brain regions aim to directly assess mental load but are difficult to collect in most laboratory IR studies let alone in natural settings. Self report instruments (e.g., [12]) are sometimes used and we have used such scales in studies of a distributed video tool [10]. Another approach is to use secondary tasks (e.g., scan for random visual or audio signals) as adjuncts to the primary search tasks. Bruza et al., [2] used secondary task monitoring in a study of interactive search in the web environment, but this technique has not yet been applied to video retrieval.

The Open Video Proejct has developed two classes of cognitive measures: recognition and inference and these measures are described in section 5.

Human attitudes and beliefs are even more difficult to assess as they are dependent primarily on self report. In our video retrieval work, we have mainly limited our evaluations to measures of user satisfaction as instantiated by scales for learnability and usability and by scales for levels of engagement and enjoyment. Such self report measures are inherently problematic as people have little basis for comparison and often overestimate their own abilities. Nonetheless, measures of satisfaction and other attitudinal measures do provide another set of indicators for human search performance and should be included along with cognitive and physical measures.

5. OPEN VIDEO PERFORMANCE MEASURES

One of the aims of the Open Video Project research is to create non-textual surrogates that enhance the search experience. Much of the evaluation is thus devoted to assessing the effectiveness of these surrogates rather than the overall search process as embedded in authentic search needs. Our approach has thus combined the traditional IR approach of abstracting and focusing while adopting laboratory studies of human performance as practiced in the human-computer interaction and usability communities. Over the course of several laboratory studies, a suite of six types of cognitive performance measures for the effectiveness of visual surrogates were developed and applied. The individual measures address recognition and inference as two cognitive activities. The three kinds of recognition measures are object recognition with textual stimuli, object recognition with visual stimuli, and action recognition with moving visual stimuli. The three kinds of inference measures are gist determination expressed as free text, gist determination selected from multiple gist statements, and visual gist determination expressed as inferences about visual stimuli.

Our general procedure is to conduct laboratory studies where, subjects are exposed to a surrogate (e.g., a fast forward at a specific rate; a storyboard with specific number of keyframes for a specific amount of time) and then asked to complete one of the performance tasks. In most cases, a five-point Likert scale for confidence is also included in the experimental protocol. Over the course of a dozen user studies, variations of these performance measures were used to assess the efficacy of surrogate designs and also to refine and validate these measures. For theoretical rationale building on image and film theory and details of these measures for two specific user studies, see the technical report [19] and for a summary of a dozen user studies and a human-centered model of effort-outcomes for video retrieval, see [9]. Additionally, measures of user satisfaction and flow have been adapted for use in assessing the affective perceptions of people using these surrogates.

5.1 Recognition Measures

For the recognition tasks, subjects view a surrogate for a specific amount of time, the surrogate is removed and the subject's recall about what they have seen is examined. Object recognition with textual stimulus provides a list of words (e.g., airplane, tree) and subjects are asked to select those words that describe objects seen in the surrogate. Twelve terms are used, half represented in the surrogate and half not in the surrogate, and in each of these two sets half are concrete terms (e.g., fountain) and half are abstract terms (e.g., archaeology). Scoring is simple and precise and scores have been shown to depend on the quality of the surrogate. For example, in fast forwards, scores are monotonically related to the rate of display.

Object recognition with visual stimuli provides a set of 12 keyframes and subjects are asked to select which keyframes were included in the surrogate. Six of the keyframes are selected from the surrogate, three from a different video with similar style, and three from a video with a different style. As with the word lists, scoring is simple and precise, although this measure assumes that the surrogates were keyframe-based surrogates. Figure 1 illustrates the on-screen task used in one study. Although subjects

are easily able to perform well on both kinds of object recognition tasks, correlations between the measures in two studies were not high, thus suggesting that they are measuring distinct kinds of recognition [20].

Action recognition is less direct than the other two forms of recognition as it provides short clips (2-3 seconds) and asks subjects to state whether they came from the video that contained the surrogate. Six clips are used: two from the segments that yielded the surrogate, two from segments similar to the video that yielded the surrogate, and two from unrelated videos. Although performance did decrease as surrogate quality decreased (e.g., faster fast forwards), action recognition was not correlated with the other two object recognition measures. Although we aimed to used learning style (visual vs verbal) as a covariate in one study to determine whether the recognition measures are dependent on personal ability/preference, we did not have enough variance in our subject pool to make a determination. This is a thread of work that bears future investigation.

Recognition is important as an antecedent to understanding and these measures give us a sense of the adequacy of surrogates, which presumably are related to the overall retrieval experience. In current work, audio surrogates are under investigation and spoken descriptions will be substituted for the keyframe-based surrogates and performance task stimuli.

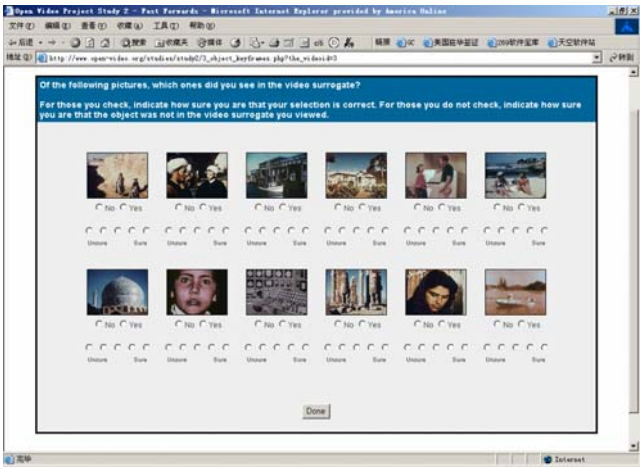

Figure 1. Object Recognition Task

5.2 Inference Measures

Gist is the sense of what some information is 'about.' A good video surrogate presumably helps the viewer determine the gist of the full video. The most direct way to determine gist is to ask people to write or say what they think the video is about after viewing the surrogate, thus inferring meaning for the video from the surrogate. This is the free text gist determination task where people are asked to write their responses in free form. In our studies, most subjects tended to write one or two word descriptions, although some subjects wrote multiple sentences. Free form expression has the advantage of directness but the disadvantage of being difficult to precisely scale. In our early work, we used two judges to score the statements on a four point scale: not correct, contains literal objects or events depicted, contains general thematic information, and contains accurate thematic information. The inter-rater agreements were quite varied across different videos, so we began using two separate two-dimensional scales: one for objects and events with accuracy and detail scores, and one for theme or plot with accuracy and detail scores. These scales yielded much more reliable inter-rater agreements.

An easier way to obtain and score gist is to provide stimulus statements and have subjects select the best one. This has obvious advantages for administration and also for scoring and we have used this multiple-choice measure in conjunction with the free text measure. After seeing the surrogate a list of sentence-length summaries are displayed and the subject selects the best summary for the video—thus making an inference about the overall video content based upon the surrogate alone. Subjects in our studies performed better on the multiple choice measures and there were only weak correlations among the two linguistic gist measures. It is important that if both kinds of gist determination measures are used that the free text is used first as the multiple choice statements offer powerful learning effects and will strongly influence what people write in free form if they have already read and completed the multiple choice stimuli.

Gist is a linguistic concept and we aim to include visual senses of aboutness. We define visual gist (vist) as the visual identity or aboutness of a surrogate or video segment. To measure vist, after viewing a surrogate or segment, a set of keyframes are shown and subjects are asked to indicate which ones 'belong' to the video from which the surrogate or segment. Of the 12 keyframes, six were selected from the video but not seen in a surrogate, three were selected from a video with similar style, and three from a video with a different style. In our studies, people were able to easily understand and complete the vist estimates accurately. As was the case with the recognition measures, people perform better (make fewer errors) when the video styles are quite distinct. We found weak correlations with the linguistic gist scores, thus indicating that vist is measuring something different than gist. Analysis of interview data over different studies lead us to believe that vist is a combination of topicality, narrative structure, and visual style (e.g., amount of motion, color, animation/real, etc.). It is my estimation that finding ways to measure vist is one of the most important research challenges because it measures unique aspects of the medium and thus has few precedents we may build upon.

5.3 Satisfaction

In addition to the cognitive measures, we assess people's perceptions about surrogates. We have adapted from the Questionnaire for User Interface Satisfaction [3], user satisfaction measures for usability and learnability. A set of six Likert-scaled statements are used to assess usability (e.g., This system makes it easier to find information) and six to assess learnabilty (e.g., learning to operate this system was easy for me). Additionally, we have adapted seven point semantic differential scales for engagement (e.g., I felt: absorbed intensely—not absorbed intensely) and enjoyment (e.g., using the system was: interesting—uninteresting). Although these measures are not direct measures of performance, they do provide indications about possible adoption of system features and may influence performance over long-time use in natural settings.

6. CONCLUSION

Measuring human performance during search is mainly subjective and leads to much less precise results than measuring system performance. It is nonetheless crucial to understanding why systems succeed or fail and in developing new kinds of system features that will improve overall video search episodes. The Open Video Digital Library project team has developed several measures for assessing the efficacy of visual surrogates and these measures have proven helpful over time in improving the features and interface of the system. Current efforts aim to continue to develop new kinds of surrogates, especially audio surrogates, and we will use and expand upon these measures in future evaluations. We also are investigating biometric measures as adjuncts to the cognitive measures so that we will have sets of measures for all three classes of human measures (physical, cognitive, and affective). It is important to note that these measures address different aspects of the search process and human interaction with retrieval systems. Thus, rather than expecting a set of highly correlated measures that might be aggregated into a single measure of search performance, we believe that the complexities of human-video retrieval system interaction are such that suites of measures will be necessary to understand the overall effects of search episodes and how those search episodes fit into people's larger goals and activities.

ACKNOWLEDGMENTS

This work has been supported by NSF grants # IIS 0099638, and IIS 0455970. Barbara Wildemuth, Gary Geisler, and Meng Yang and other members of the Open Video Project team developed and applied these measures.

7. REFERENCES

[1] Boekelheide, K., Brown, A., Fu, X., Marchionini, G., Oh, S., Rogers, G., Saelim, B., Song, Y., & Stutzman, F. (2006). Audio surrogation for digital video: A design framework. SILS Technical Report TR-2006-02. http://sils.unc.edu/research/publications/reports/TR-2006-02.pdf

[2] Bruza, P., McArthur, R., & Dennis, S. (2000). Interactive internet search: Keyword, directory, and query reformulation mechanisms compared. *Proc. of SIGIR 2000.* (Athens, Greece, July 24-28, 2000), 280-287.

[3] Chin, JP, Diehl, VA, and Norman, KL. Development of an instrument measuring user satisfaction of the human-computer interface. *Proceedings of SIGCHI '88.* 213-218. 88. New York, ACM/SIGCHI.

[4] Fluhr, C., Moellic, P., & Hede, P. (in press). Usage-oriented multimedia information retrieval technological evaluation. Multimedia Information Retrieval Workshop (Santa Barbara, CA, October 26027, 2006). NY: ACM Press

[5] Hughes, A., Wilkens, T., Wildemuth, B., & Marchionini, G. (2003). Text or pictures? An eyetracking study of how people view digital video surrogates. *Proceedings of the International Conference on Image and Video Retrieval (CIVR 2003),* 271-280.

[6] Marchionini, G. (2006). Exploratory Search. Comm. Of the ACM. Fröhlich, B. and Plate, J. The cubic mouse: a new device for three-dimensional iput. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI '00)* (The Hague, The Netherlands, April 1-6, 2000). ACM Press, New York, NY, 2000, 526-531.

[7] Li, Y., Dorai, C., & Farrell, R. (2005). Creating MAGIC: System for generating learning object metadata for instructional content. *Proceedings of ACM Multimedia 2005* (Singapore, Nov. 6-11, 2005).

[8] Marchionini, G (2006). Exploratory search: From finding to understanding. *Communications of the ACM*, 49(4), p. 41-46.

[9] Marchionini, G., Wildemuth, B., & Geisler, G. (in press). The Open Video Digital Library: A Mobius strip of research and practice. *Journal of the American Society for Information Science & Technology*.

[10] Mu, X. (2004). *SmartLinks in a video-based collaborative distance learning system: A cognitive model and evaluation study*. Unpublished doctoral dissertation, University of North Carolina at Chapel Hill.

[11] Over, P., Kraaij, W., & Smeaton, A. (2005). TRECVID 2005: An introduction. *Proc. TRECVID 2005* (Gaithersburg, MD), 1-14. http://www.cdvp.dcu.ie/Papers/TRECVid2005_Overview.pdf

[12] Reid, G.B. & Colle, H.A. (1988). Critical SWAT values for predicting operator overload. In *Proceedings of the Human Factors Society 32nd annual meeting*. Santa Monica, CA: Human Factors Society, 1414-1418.

[13] Smeaton, A., Over, P., & Kraaij, W. (in press). Evaluation campaigns and TRECVid. Multimedia Information Retrieval Workshop (Santa Barbara, CA, October 26027, 2006). NY: ACM Press

[14] Sparck Jones, K. (2006). What's the value of TREC—is there a gap to jump or a chasm to bridge? *SIGIR Forum*, 40(1), 10-20.

[15] Wactlar, H., Christel, M., Gong, Y., & Hauptmann, A. (1999). Lessons learned from building a terabyte digital video library. *Computer*, *32*(2), 66-73.

[16] Westerveld, T. & van Zwol, R. (in press). Benchmarking multimedia search in structured collections. Multimedia Information Retrieval Workshop (Santa Barbara, CA, October 26027, 2006). NY: ACM Press

[17] White, R. & Marchionini, G. (in press). Examining the effectiveness of real-time query expansion. *Information Processing & Management.*

[18] White, R., Muresan, G., & Marchionini, G. (in press). Evaluating exploratory search systems. *SIGIR Workshop* paper (Seattle, August 7-11, 2006).

[19] Yang, M., Wildemuth, B. M., Marchionini, G., Wilkens, T., Geisler, G., Hughes, A., Gruss, R., & Webster, C. (2003b). *Measures of User Performance in Video Retrieval Research. SILS Technical Report 2003-02.* Chapel Hill, NC: University of North Carolina, School of Information and Library Science. http://www.ils.unc.edu/ils/research/TR-2003-02.pdf.

[20] Yang, M., Wildemuth, B., Marchionini, G., Wilkens, T., Geisler, G., Hughes, A., Gruss, R., & Webster, C. (2003). Measuring user performance during interactions with digital video collections. *Proc of the American Society for Information Science & Technology*, (Long Beach, CA, Oct. 19-23, 2003), Silver Spring, MD: ASIST. 3-11.