International Conference on Image and Video Retrieval (CIVR 2006), July 13-15 ,2006, Tempe, AZ

Mining Novice User Activity with TRECVID Interactive Retrieval Tasks

Michael G. Christel and Ronald M. Conescu

School of Computer Science, Carnegie Mellon University
Pittsburgh, PA, U.S.A. 15213

christel@cs.cmu.edu, rconescu@andrew.cmu.edu

Abstract. This paper investigates the applicability of Informedia shot-based interface features for video retrieval in the hands of novice users, noted in past work as being too reliant on text search. The Informedia interface was redesigned to better promote the availability of additional video access mechanisms, and tested with TRECVID 2005 interactive search tasks. A transaction log analysis from 24 novice users shows a dramatic increase in the use of color search and shot-browsing mechanisms beyond traditional text search. In addition, a within-subjects study examined the employment of user activity mining to suppress shots previously seen. This strategy did not have the expected positive effect on performance. User activity mining and shot suppression did produce a broader shot space to be explored and resulted in more unique answer shots being discovered. Implications for shot suppression in video retrieval information exploration interfaces are discussed.

1 Introduction

As digital video becomes easier to create and cheaper to store, and as automated video processing techniques improve, a wealth of video materials are now available to end users. Concept-based strategies, where annotators carefully describe digital video with text concepts that can later be used for searching and browsing, are powerful but expensive. Users have shown that they are unlikely to invest the time and labor to annotate their own photograph and video collections with text descriptors. Prior evaluations have shown that annotators do not often agree on the concepts used to describe the materials, so the text descriptors are often incomplete.

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

Continuing research in the video information indexing and retrieval community attempts to address the semantic gap by automatically deriving higher order features, e.g., outdoor, building, face, crowd, and waterfront. Rather than leave the user only

with color, texture, and shape, these strategies give the user control over these higher order features for searching through vast corpora of materials. The NIST TRECVID video retrieval evaluation forum has provided a common benchmark for evaluating such work, charting the contributions offered by automated content-based processing as it advances [1].

To date, TRECVID has confirmed that the best performing interactive systems for news and documentary video leverage heavily from the narration offered in the audio track. The narration is transcribed either in advance for closed-captioning by broadcasters, or as a processing step through automatic speech recognition (ASR). In this manner, text concepts for concept-based retrieval are provided for video, without the additional labor of annotation from a human viewer watching the video, with the caveat that the narration does not always describe the visual material present in the video. Because the text from narration is not as accurate as a human annotator describing the visual materials, and because the latter is too expensive to routinely produce, subsequent user search against the video corpus will be imprecise, returning extra irrelevant information, and incomplete, missing some relevant materials as some video may not have narrative audio. Interfaces can help the interactive user to quickly and accurately weed out the irrelevant information and focus attention on the relevant material, addressing precision. TRECVID provides an evaluation forum for determining the effectiveness of different interface strategies for interactive video retrieval. This paper reports on a study looking at two interface characteristics:

1. Will a redesigned interface promote other video information access mechanisms besides the often-used text search for novice users?
2. Will mining user interactions, to suppress the future display of shots already seen, allow novice users to find more relevant video footage than otherwise?

The emphasis is on novice users: people who are not affiliated with the research team and have not seen or used the system before. Novices were recruited as subjects for this experiment to support the generalization of experimental results to wider audiences than just the research team itself.

3 Informedia Retrieval Interface for TRECVID 2005

The Informedia interface since 2003 has supported text query, image color-based or texture-based query, and browsing actions of pre-built "best" sets like "best road shots" to produce sets of shots and video story segments for subsequent user action, with the segments and shots represented with thumbnail imagery often in temporally arranged storyboard layouts [2, 3, 4]. Other researchers have likewise found success with thumbnail image layouts confirmed with TRECVID studies [5, 6, 7]. This paper addresses two questions suggested by earlier TRECVID studies.

First, Informedia TRECVID 2003 experiments suggested that the usage context from the user's interactive session could improve one problem with storyboard interfaces on significantly sized corpora: there are too many shot thumbnails within candidate storyboards for the user's efficient review. The suggestion was to mark all shots seen by a user pursuing a topic, and suppress those shots from display in subsequent interactions regarding the topic [4]. Mining the users' activity in real time can reduce the number of shots shown in subsequent interactions.

Second, Informedia TRECVID 2004 experiments found that novice users do not pursue the same solution strategies as experts, using text query for 95% of their investigations even though the experts' strategy made use of image query and "best" set browsing 20% of the time [3]. The Informedia interface for TRECVID 2005 was redesigned with the same functionality as used in 2003 and 2004, but with the goal of promoting text searches, image searches, and visual feature browsing equally. Nielsen's usability heuristics [8] regarding "visibility of system status" and "recognition over recall," and guidelines for clarifying search in text-based systems [9] were consulted during the updating, with the redesigned Informedia interface as used for TRECVID 2005 shown in Fig. 1.

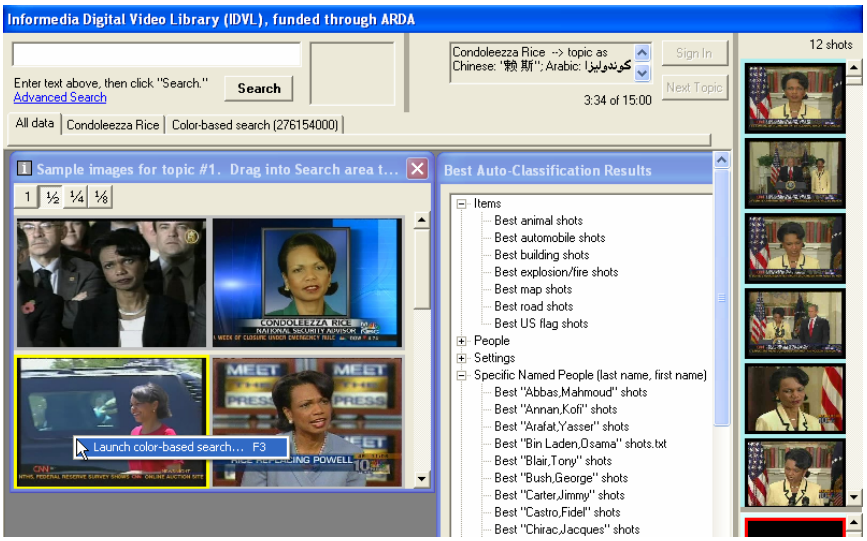

Fig. 1. 2005 Informedia TRECVID search interface, with text query (top left), image query (middle left), topic description (top middle), best-set browsing (middle), and collected answer set display (right) all equally accessible

Fig. 2 illustrates a consistency in action regarding the thumbnail representations of shots. The shot can be captured (saved as an answer for the topic), used to launch an image query, used to launch a video player queued to that shot's start time, used to launch a storyboard of all the shots in the shot's story segment, or used to show production metadata (e.g., date of broadcast) for the segment. New for 2005 was the introduction of 2 capture lists supporting a ranked answer set: the "definite" shots in the "yes" list, and the "possible" answers put to a "maybe" secondary list. The six actions were clearly labeled on the keyboard by their corresponding function keys.

As an independent variable, the interface was set up with the option to aggressively hide all previously "judged" shots. While working on a topic, the shots seen by the user as thumbnails were tracked in a log. If the user captured the shot to either the "yes" or "maybe" list, it would not be shown again in subsequent text and image queries, or subsequent "best" set browsing, as these shots were already judged posi-

tively for the topic. In addition, all shots skipped over within a storyboard while capturing a shot were assumed to be implicitly judged negatively for the topic, and would not be shown again in subsequent user actions on that topic. So, for the topic of "tanks or military vehicles", users might issue a text search "tank" and see a storyboard of shots as in Fig. 2. They capture the third shot on the top row. That shot, and the first 2 shots in that row marked as "implicitly judged negatively", are now no longer shown in subsequent views. Even if those 3 shots discuss "soldier", a subsequent search on "soldier" would not show the shots again. The "implicitly judged negatively" shots, henceforth labeled as "overlooked" shots, are not considered further, based on the assumption that if a shot was not good enough for the user to even capture as "maybe", then it should not be treated as a potentially relevant shot for the topic.

Fig. 2. Context-sensitive menu of shot-based actions available for all thumbnail representations in the Informedia TRECVID 2005 interface

3 Participants and Procedure

Study participants were recruited through electronic communication at Carnegie Mellon University. The 24 subjects who participated in this study had no prior experience with the interface or data under study and no connection with the TRECVID research group. The subjects were 11 female and 13 male with a mean age of 24 (6 subjects less than 20, 4 30 or older); 9 undergraduate students, 14 graduate students, and 1 university researcher. The participants were generally familiar with TV news. On a 5-point scale responding to "How often do you watch TV news?" (1=*not at all*, 5=*more than once a day*), most indicated some familiarity (distribution for 1-5 were 6-3-8-5-2). The participants were experienced web searchers but inexperienced digital video searchers. For "Do you search the web/information systems frequently?" (1=*not at all*, 5= *I search several times daily*), the answer distribution was 0-1-3-7-13 while for "Do you use any digital video retrieval system?" with the same scale, the distribution was 15-7-1-1-0. These characteristics are very similar to those of novice users in a TRECVID 2004 published study [3]. Each subject spent about 90 minutes in the study and received $15 for participation.

Participants worked individually with an Intel® Pentium® 4 class machine, a high resolution 1600 x 1200 pixel 21-inch color monitor, and headphones in a Carnegie Mellon computer lab. Participants' keystrokes and mouse actions were logged within the retrieval system during the session. They first signed a consent form and filled out

a questionnaire about their experience and background. During each session, the participant was presented with four topics, the first two presented with one system ("Mining" or "Plain") and the next two with the other system. The "Mining" system kept track of all captured shots and overlooked shots. Captured and overlooked shots were not considered again in subsequent storyboard displays, and overlooked shots were skipped in filling out the 1000-shot answer set for a user's graded TRECVID submission. The "Plain" system did not keep track of overlooked shots and did not suppress shots in any way based on prior interactions for a given topic.

24 subject sessions produced 96 topic answers: 2 Mining and 2 Plain for each of the 24 TRECVID 2005 search topics. The topics and systems were counter-balanced in this within-subjects design: half the subjects experienced Mining first for 2 topics, and then Plain on 2 topics, while the other half saw Plain first and then Mining. For each topic, the user spent exactly 15 minutes with the system answering the topic, followed by a questionnaire. The questionnaire content was the same as used in 2004 across all of the TRECVID 2004 interactive search participants, designed based on prior work conducted as part of the TREC Interactive track for several years [10]. Participants took two additional post-system questionnaires after the second and fourth topics, and finished with a post-search questionnaire.

Participants were given a paper-based tutorial explaining the features of the system with details on the six actions in Figure 2, and 15 minutes of hands-on use to explore the system and try the examples given in the tutorial, before starting on the first topic.

4 Results

Transaction log analysis shows that the interface changes illustrated in Fig. 1 had their intended effect: novice users made much more frequent use of image search and the "best" sets browsing rather than relying almost exclusively on text search, as was found in the TRECVID 2004 experiment [3]. Table 1 summarizes the results, including an extra column showing the statistics from the TRECVID 2004 experiment. Since the topics, corpus, and features for "best" sets changed between TRECVID 2004 and 2005, the reader is cautioned against making too many inferences between corpora. For example, the increase in segment count per text query from 2004 to 2005 might be due to more ambiguous queries of slightly longer size, but could also be due to the TRECVID 2005 overall corpus being larger. The point emphasized here is that with the 2004 interface, novices were reluctant to interact with the Informedia system aside from text search, while in 2005 the use of "best" set browsing increased ten-fold and image queries three-fold.

Table 2 shows the access mechanism used to capture shots and the distribution of captured correct shots as graded against the NIST TRECVID pooled truth [1]. While "best" browsing took place much less than image search (see Table 1), it was a more precise source of information, producing a bit more of the captured shots than image search and an even greater percentage of the correct shots. This paper focuses on novice user performance; the expert's performance is listed in Table 2 only to show the relative effects of interface changes on novice search behavior compared to the expert. In 2004, the expert relied on text query for 78% of his correct shots, with image query shots contributing 16% and "best" set browsing 6%. The novice users' interactions were far different, with 95% of the correct shots coming from text search

and near nothing coming from "best" set browsing. By contrast, the same expert in 2005 for the TRECVID 2005 topics and corpus drew 53% of his correct shots from text search, 16% from image query, and 31% from "best" set browsing. The novice users with the same "Mining" interface as used by the expert produced much more similar interaction patterns to the expert than was seen in 2004, with image query and "best" set browsing now accounting for 35% of the sources of correct shots.

Table 1. Interaction log statistics for novice user runs with TRECVID data

	TRECVID 2005		*TRECVID 2004 Novice ([3])*
	Novice Plain	Novice Mining	
Number of users	24	24	*24*
Number of topics	48	48	*48*
Fixed minutes spent per topic	15	15	*15*
Avg. (average) feature "best" sets browsed per topic	1.38	1.13	*0.13*
Avg. image queries per topic	3.27	4.19	*1.23*
Avg. text queries per topic	5.67	7.21	*9.04*
Word count per text query	2.31	2.19	*1.51*
Avg. number of video segments returned by each text query	194.7	196.8	*105.3*
Query/browse actions per topic	10.32	12.53	*10.4*

Table 2. Percentages of submitted shots and correct shots from various groups

	Access Mechanism	TRECVID 2005			*TRECVID 2004 ([3])*	
		Novice Plain	Novice Mining	Expert Mining	*Expert*	*Novice*
Shots Submitted	Text query	48%	65%	54%	*81%*	*95%*
	Image query	25%	17%	20%	*12%*	*5%*
	"Best" browse	27%	18%	26%	*7%*	*0%*
Shots Judged Correct	Text query	47%	65%	53%	*78%*	*95%*
	Image query	24%	14%	16%	*16%*	*5%*
	"Best" browse	29%	21%	31%	*6%*	*0%*

Overall performance for the novice runs was very positive, with the mean average precision (MAP) for four novice runs of 0.253 to 0.286 placing the runs in the middle of the 44 interactive runs for TRECVID 2005, with all of the higher scoring runs coming from experts (developers and colleagues acting as users of the tested systems). Hence, these subjects produced the top-scoring novice runs, with the within-subjects study facilitating the comparison of one system vs. another, specifically the relative merits of Plain vs. Mining based on the 96 topics answered by these 24 novice users.

There is no significant difference in performance as measured by MAP for the 2 Plain and 2 Mining runs: they are all essentially the same. Mining did not produce

the effect we expected, that suppressing shots would lead to better average precision for a topic within the 15-minute time limit. The users overwhelmingly (18 of 24) answered "no difference" to the concluding question "Which of the two systems did you like best?" confirming that the difference between Plain and Mining was subtle (in deciding *what* to present) rather than overt in *how* presentation occurs in the GUI. The Mining interface did lead to more query and browsing interactions, as shown in the final row of Table 1, and while these additional interactions did not produce an overall better MAP, they did produce coverage changes as discussed below.

5 Discussion

TRECVID encourages research in information retrieval specifically from digital video by providing a large video test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results. TRECVID benchmarking covers interactive search, and the NIST TRECVID organizers are clearly cognizant of issues of 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. Regarding the task context, TRECVID organizers design interactive retrieval topics to reflect many of the various sorts of queries real users pose [1]. Regarding the user pool, if only the developers of the system under study serve as the users, it becomes difficult to generalize that novices (non-developers and people outside of the immediate research group) would have the same experiences and performance. In fact, a study of novice and expert use of the Informedia system against TRECVID 2004 tasks shows that novice search behavior is indeed different from the experts [3]. Hence, for TRECVID user studies to achieve greater significance and validity, they should be conducted with user pools drawn from communities outside of the TRECVID research group, as done for the study reported here, its predecessor novice study reported in [3], and done with [11].

The interface design can clearly affect novice user interaction. A poor interface can deflate the use of potentially valuable interface mechanisms, while informing the user as to what search variants are possible and all aspects of the search (in line with [9]) and promoting "visibility of system status" and "recognition over recall" [8] can produce a richer, more profitable set of user interactions. The Informedia TRECVID 2005 interface succeeded in promoting the use of "best" browsing and image search nearly to the levels of success achieved by an expert user, closing the gulf between novice and expert interactions witnessed with a TRECVID 2004 experiment.

As for the Mining interface, it failed to produce MAP performance improvements. The TRECVID 2005 interactive search task is specified to allow the user 15 minutes on each of 24 topics to identify up to 1000 relevant shots per topic from among the 45,765 shots in the test corpus. As discussed in [5], MAP does not reward returning a short high precision list over returning the same list supplemented with random choices, so the search system is well advised to return candidate shots above and beyond what the user identifies explicitly. For our novice runs, the "yes" primary captured shot set was ranked first, followed by the "maybe" secondary set of captured shots (recall Fig. 2 options), followed by an automatic expansion seeded by the user's captured set, to produce a 1000 shot answer set. For the Mining treatment, the over-

looked shots were never brought into the 1000 shot answer set during the final automatic expansion step. A post hoc analysis of answer sets shows that this overly aggressive use of the overlooked shots for Mining was counterproductive. The user actually identified more correct shots with Mining than with Plain. Table 3 summarizes the results, with the expert run from a single expert user with the Mining system again included to illustrate differences between performances obtained with the Mining and Plain system. Novices for both named specific topics and generic topics, as classified by the TRECVID organizers [1], had better recall of correct shots in their primary sets with Mining versus Plain. However, the precision was less with Mining, perhaps because when shots are suppressed, the user's expectations are confounded and the temporal flow of thumbnails within storyboards is broken up by the removal of overlooked shots. Suppressing information has the advantage of stopping the cycle of constant rediscovery of the same information, but has the disadvantage of removing navigation cues and the interrelationships between data items [12], in our case shots. Coincidentally, the novices did use the secondary capture set as intended, for lower precision guesses or difficult-to-confirm-quickly shots: the percentage correct in the secondary set is less than the precision of the primary capture set.

Table 3. Primary and secondary captured shot set and overlooked shot set sizes, with percentage of correct shots, for named and generic TRECVID 2005 topics

TRECVID 2005	Shot Set	Avg. Shot Count Per Topic			% Correct in Shot Set		
		Novice Plain	Novice Mining	Expert Mining	Novice Plain	Novice Mining	Expert Mining
6 Named Topics	Primary	53.9	64.8	67.2	92.4	72.1	97.0
	Secondary	5.3	5.2	12.2	31.3	46.8	93.2
	Overlooked	n/a	372.0	223.7	n/a	5.8	8.6
18 Generic Topics	Primary	41.9	47.9	55.7	78.3	74.9	91.2
	Secondary	5.7	3.8	11.9	42.2	26.8	65.1
	Overlooked	n/a	649.8	503	n/a	4.2	6.5

The most glaring rows from Table 3 address the overlooked shot set (suppressed shots that are not in the primary or secondary capture sets): far from containing no information, they contain a relatively high percentage of correct shots. A random pull of shots for a named topic would have 0.39% correct shots, but the novices' overlooked set contained 5.8% correct items. A random pull of generic shots would contain 0.89% correct shots, but the novices' overlooked set contained 4.2% correct shots. Clearly, the novices (and the expert) were overlooking correct shots at a rate higher than expected.

Fig. 3 shows samples of correct shots that were overlooked when pursuing the topic "Condoleezza Rice." They can be categorized into four error types: (a) the shot was mostly of different material that ends up as the thumbnail representation, but started or ended with a tiny bit of the "correct" answer, e.g., the end of a dissolve out of a Rice shot into an anchor studio shot; (b) an easily recognizable correct shot based on its thumbnail, but missed by the user because of time pressure, lower motivation

than the expert "developer" users often employed in TRECVID runs, and lack of time to do explicit denial of shots with "implicitly judged negatively" used instead to perhaps too quickly banish a shot into the overlooked set; (c) incorrect interpretation of the query (Informedia instructions were to ignore all still image representations and only return video sequences for the topic); and (d) a correct shot but with ambiguous or incomplete visual evidence, e.g., back of head or very small. Of these error classes, (a) is the most frequent and easiest to account for: temporal neighbors of correct shots are likely to be correct because relevant shots often occur in clumps and the reference shot set may not have exact boundaries. Bracketing user-identified shots with their neighbors during the auto-filling to 1000 items has been found to improve MAP by us and other TRECVID researchers [5, 7]. However, temporally associated shots are very likely to be shown in initial storyboards based on the Informedia storyboard composition process, which then makes neighbor shots to correct shots highly likely to be passed over, implicitly judged negatively, and, most critically, never considered again during the auto-filling to 1000 shots. So, the aggressive mining and overlooking of shots discussed here led to many correct shots of type (a) being passed over permanently, where bracketing strategies as discussed in [5] would have brought those correct shots back into the final set of 1000.

Fig. 3. Sample of overlooked but correct shots for Condoleezza Rice topic, divided into 4 error classes (a) - (d) described above

One final post hoc analysis follows the lines of TRECVID workshop inquiries into unique relevant shots contributed by a run. Using just the 4 novice runs and one expert run per topic from the Informedia interactive system, the average unique relevant shots in the primary capture set contributed by the novices with Plain was 5.1 per topic, novices with Mining contributed 7.1, and the expert with Mining for reference contributed 14.9. Clearly the expert is exploring video territory not covered by the novices, but the novices with the Mining interfaces are also exploring a broader shot space with more unique answer shots being discovered.

6 Summary and Acknowledgements

Video retrieval achieves higher rates of success with a human user in the loop, with the interface playing a pivotal role in encouraging use of different access mechanisms by novices. A redesigned Informedia interface succeeded in promoting the use of

image search and "best" shot set browsing in addition to text search, as evidenced by 24 novice user sessions addressing 4 TRECVID 2005 topics each. These sessions also served as a within-subjects experiment to test whether an aggressive strategy for hiding previously captured and passed over shots would produce better performance. The Mining interface with such shot suppression did not perform better than the control Plain interface using the TRECVID metric of MAP. However, the Mining strategy appears promising where diversity in answer sets is rewarded, e.g., if finding 3 answer shots from 3 different sources on different reporting days is more important than finding a clump of 3 answer shots temporally adjacent to one another. Also, by relaxing the aggressive restriction of overlooked shots, the Mining strategy can be adjusted to still encourage more diverse exploration by the user, but then to recover suppressed shots that are temporal neighbors to answer shots (as shown in Fig. 3a). Our revised Mining strategy will still suppress overlooked shots during the user interaction period, but then will not ignore the overlooked shots during the auto-filling to the complete (1000) TRECVID answer set, reverting back to the temporal bracketing [5] that has proven useful in the past.

This material was made possible by the NIST assessors and the TRECVID community. It is based on work supported by the Advanced Research and Development Activity (ARDA) under contract number H98230-04-C-0406 and NBCHC040037, and supported by the National Science Foundation under Grant No. IIS-0535056.

References

1. Over, P., Ianeva, T., Kraaij, W., Smeaton, A.F.: TRECVID 2005 An Introduction. TRECVID 2005 Proceedings, http://www-nlpir.nist.gov/projects/trecvid
2. Hauptmann, A.G.: Lessons for the Future from a Decade of Informedia Video Analysis Research. Proc. CIVR (Singapore, July 2005), LNCS 3568: 1-10
3. Christel, M., Conescu, R.: Addressing the Challenge of Visual Information Access from Digital Image and Video Libraries. Proc. ACM/IEEE JCDL (Denver, June 2005), ACM Press, 69-78
4. Christel, M., Moraveji, N.: Finding the Right Shots: Assessing Usability and Performance of a Digital Video Library Interface. Proc. ACM Multimedia (New York, Oct. 2004), ACM Press, 732–739
5. Adcock, J., Cooper, M., Girgensohn, A., Wilcox, L.: Interactive Video Search Using Multilevel Indexing. Proc. CIVR (Singapore, July 2005), LNCS 3568: 205-214
6. Snoek, C., Worring, M., et al.: MediaMill: Exploring News Video Archives based on Learned Semantics. Proc. ACM Multimedia (Singapore, Nov. 2005), ACM Press, 225-226.
7. Hauptmann, A., Christel, M.: Successful Approaches in the TREC Video Retrieval Evaluations. Proc. ACM Multimedia (New York, Oct. 2004), ACM Press, 668-675
8. Nielsen, J.: Heuristic Evaluation. In Nielsen, J., and Mack, R.L. (eds.), Usability Inspection Methods. John Wiley & Sons, New York, NY, 1994
9. Shneiderman, B., Byrd, D., Croft, W.B.: Clarifying Search: A User-Interface Framework for Text Searches. D-Lib Magazine, 3, 1 (January 1997), http://www.dlib.org
10. Kraaij, W., Smeaton, A.F., Over, P., Arlandis, J.: TRECVID 2004 – An Introduction. In TRECVID'04 Proc., http://www-nlpir.nist.gov/projects/tvpubs/tvpapers04/tv4overview.pdf
11. Yang, M., Wildemuth, B., Marchionini, G.: The Relative Effectiveness of Concept-based Versus Content-based Video Retrieval. Proc. ACM Multimedia 2004, ACM Press 368-371
12. Golovchinsky, G.: Queries? Links? Is there a difference? Proc. CHI '97, ACM Press (1997), 407-414