Merging Storyboard Strategies and Automatic Retrieval for Improving Interactive Video Search

Michael G. Christel
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213 USA
1-412-268-7799
christel@cs.cmu.edu

Rong Yan
IBM TJ Watson Research Center
Hawthorne, NY 10532 USA
1-914-784-7142
yanr@us.ibm.com

ABSTRACT
The Carnegie Mellon University Informedia group has enjoyed consistent success with TRECVID interactive search using traditional storyboard interfaces for shot-based retrieval. For TRECVID 2006 the output of automatic search was included for the first time with storyboards, both as an option for an interactive user and in a different run as the sole means of access. The automatic search makes use of relevance-based probabilistic retrieval models to determine weights for combining retrieval sources when addressing a given topic. Storyboard-based access using automatic search output outperformed extreme video retrieval interfaces of manual browsing with resizable pages and rapid serial visualization of keyframes that used the same output. Further, the full Informedia interface with automatic search results as an option along with other query mechanisms scored significantly better than all other TRECVID 2006 interactive search systems. Attributes of the automatic search and interactive search systems are discussed to further optimize shot-based retrieval from news corpora.

Categories and Subject Descriptors
H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems – *evaluation, video*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms
Experimentation, Human Factors

Keywords
Video retrieval, TRECVID results analysis

1. INTRODUCTION
This paper reports on the Carnegie Mellon University Informedia group's TRECVID 2006 interactive search experiences and successes, particularly the use of traditional storyboard interfaces for shot-based retrieval. In 2006 the ranked shot lists from automatic search were also provided to users, leading to a significant boost in performance. An introduction to TRECVID is given, followed by the motivation for the experiments reported here. Specifically, the new work here is as follows, conducted through a series of TRECVID 2006 search runs:

1. Can the results of fully automatic search help the user in interactive search?
2. Does the method of presentation matter, i.e., is there a difference between extreme "push many images before the user's eyes" under automatic control, versus image layout and paging in traditional storyboards under user control?
3. Does the freedom to solicit different shot sets matter, i.e., rather than being locked in to the output of a fully automatic search with interactive filtering, let the user also generate other shot sets for inspection through querying by text, imagery, and concepts?

TRECVID at NIST has 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 54 companies and academic institutions from 2001 to 2006 [6]. A number of tasks are defined in TRECVID, including shot detection, story segmentation, semantic feature extraction, and information retrieval, with this paper focused only on the retrieval task.

TRECVID video searches deal with 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. Success is measured based on quantities of relevant shots 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 shot 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. Search types include "automatic" in which the query topic is taken as is with no human modifications, 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.

CIVR'07, July 9-11, 2007, Amsterdam, The Netherlands.

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 [6]. 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. 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 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.

In each and every TRECVID prior to 2006, the Informedia storyboard-based interface [2, 3] has either produced the top score for interactive search, or produced high scores statistically clustered with the top-scoring MediaMill interface [8, 9]. It has been an interface with consistently proven effectiveness over the years, but was challenged in TRECVID 2005 with a different type of interface, an "extreme" video retrieval (XVR) interface based on rapid serial visual presentation of keyframes [5]. The XVR interface, the Informedia storyboard interface, and the MediaMill interface collectively produced significantly better interactive search runs than all other submissions in TRECVID 2005. MediaMill and Informedia shared 3 common means of querying the news corpus to produce shot sets: query-by-textual-keyword, query-by-image-example, and query-by-concept [9]. The XVR interface provided a fourth strategy: do no queries, but work only against the output of a fully automatic search result for the topic. For TRECVID 2006, we specifically examined the effects of different interfaces presenting the ranked shot output of automatic retrieval systems, to understand better the reason behind the good performance of extreme video retrieval interface. Is it because XVR is an inherently better interface than the dense storyboards used in the traditional Informedia interface, or is it due to the availability of good ranked shot output from automatic retrieval? In addition, we explored whether there would be a benefit toward including another shot set producing strategy into the Informedia storyboard interface: looking at the automatic search run output, in addition to query-by-text, query-by-image, and query-by-concept.

To facilitate easier comparison across TRECVID participants, we adopt the query access descriptions provided by Snoek and colleagues in their work [8, 9]: "query-by-textual-keyword", "query-by-image-example", and "query-by-concept." To still distinguish the primary modality of the query but to save space, we discuss these access methods as "query-by-text", "query-by-image", and "query-by-concept", respectively, with one addition: the available of the ranked shot output of the fully automatic search, which we label "query-by-best-of-topic" since this is a topic-specific shot list. The six TRECVID Informedia search runs discussed in the remainder of this paper can be summarized as follows:

- **Automatic Baseline**: automatic search based only on transcript text
- **Automatic**: automatic search encompassing all modalities text, visual, and aural
- **Informedia Full**: storyboard interface with 4 means of access: query-by-text, query-by-image, query-by-concept, and query-by-best-of-topic (the latter "query-by-best-of-topic" looks through the ranked shot list produced by the **Automatic** run)
- **Informedia Limited**: the same interface as **Informedia Full**, except that only query-by-best-of-topic is available
- **XVR-1**: the same data as **Informedia Limited**, i.e., the output of the **Automatic** run just as XVR was used with success in TRECVID 2005, but an "extreme" interface manual browsing with resizable pages
- **XVR-2**: the same data as **Informedia Limited**, but an "extreme" interface rapid serial visual presentation.

Given that all of the runs make use of automatic search, a discussion of the state of automatic search is given first, followed by the results from the different interactive search runs.

2. TRECVID AUTOMATIC SEARCH

The **Automatic** search run makes use of relevance-based probabilistic retrieval models to determine weights for combining a large set of semantic concepts when addressing a given topic [1, 4, 10]. This model translated the retrieval task into a supervised learning problem with the parameters learned discriminatively. Rather than treating retrieval as a classification problem, we used an algorithm called "ranking logistic regression" [11] by accounting for the order information between training data, so that the optimization is closely associated with the retrieval performance criteria. We also use a direct query expansion approach to automatically map the multimodal query to existing semantic concepts.

2.1 Relevance-Based Retrieval Model

Although we considered a relevance-based probabilistic retrieval model for knowledge combination in our implementation, previous work showed that simply adopting a query-independent knowledge combination strategy is not flexible enough to handle variations in users' information needs. We developed more advanced methods to incorporate the factor of query information into the probabilistic retrieval model, by making the following assumptions on the query space:

1. The entire query space can be described by a finite number of mixtures, where the queries from each mixture have the similar characteristics and share the same combination function.
2. Query descriptions can be used as indicators to which mixture the query (i.e., topic) belongs.

Based on these assumptions, the retrieval model can be represented as:

$$P(y_+ \mid D,Q) = \sum_{k=1}^{K} P(z_k \mid Q) \cdot \sigma\left(\sum_{i=0}^{N} \lambda_{ki} f_i(D,Q) \right)$$

where z_k are the variables indicating the defined query types, and $\sigma(x) = \left(1 + e^{-x}\right)^{-1}$ is the logistic function and λ_i is the combination parameters for the outputs from different knowledge sources $P(S_i|D,Q)$. In this paper, we manually defined multiple mutual exclusive query types for each query and thus limit one and only one z_k to be 1 while the other z_k set to 0. Similar to the settings employed for prior TRECVID automatic runs by our research group [10], we automatically assign each query to one of five types:

Named person: queries for finding a named person, possibly with certain actions.

Named object: queries for a specific object with a unique name or an object with consistent visual appearance.

General object: queries for a general category of objects instead of a specific one among them.

Sports: queries related to sport events.

Scene: queries depicting a scene with multiple types of objects in certain spatial relationships.

The query type classification method is based on rule-based text processing techniques, which consist of three phases: named entity extraction, tagging/chunking, and syntactic parsing [10]. Named entity extraction is mainly used to identify queries for sports, named persons and named objects, which contain proper names like people, location, or organization. Part-of-speech-tagging and noun-phrase-chunking is conducted to distinguish "general object" queries and scene queries. Finally, syntactic parsing is applied to correct the misclassified scene queries.

Input: Feature matrix $F^{M_D \times N}$ where $F_{ij} = f_i(D_j, Q)$ and $F_j = (f_{1j}, ..., f_{M_D j})$. Document relevance $Y = (Y_1, ..., Y_{M_D})^T$.
Output: Weights $\lambda = (\lambda_1, ..., \lambda_N)^T$ that maximizes the log-likelihood.
Algorithm:
Let $\lambda^{(1)}$ to be the zero vector. For $k = 1, 2, ...$

1. Compute the fitted value

$$\pi_j = \frac{\exp(F_j \lambda^{(k)})}{1 + \exp(F_j \lambda^{(k)})}, j = 1, ..., M_D$$

2. Define an $n \times n$ weight matrix W whose i^{th} diagonal element is $\pi_j(1 - \pi_j)$
3. Define the adjusted response vector

$$Z = F\lambda^{(k)} + W^{-1}(Y - \pi),$$

where $\pi = (\pi_1, ..., \pi_{M_D})$

4. Take the weighted linear regression of Z on X,

$$\lambda^{(k+1)} = (F^T W F)^{-1} F^T W Z.$$

The standard errors are given by $V(\lambda) = (F^T W F)^{-1}$.

Figure 1. The reweighted least square algorithm for the logistic regression.

After each query is associated with a single query class, the parameters can be estimated with the training data being restricted in the given query class. However, the optimal parameters have to be found numerically. Various numerical optimization algorithms, e.g., Newton's method, conjugate gradient and iterative scaling, can be applied to optimize the log-likelihood of logistic regression. In the implementation, we adopt a well-known optimization algorithm called reweighted least squares, with its learning process is shown in .

2.2 Ranking Logistic Regression

However, there are some issues if we directly cast the retrieval task into a binary classification problem such as logistic regression. For instance, in the retrieval scenario the number of positive data is much smaller than the negative data. More importantly, the optimization criterion of classification has no relationship to the retrieval performance measure, namely, the average precision. This might lead to some weird effects on the learned weights. Therefore, we proposed a new approach called "ranking logistic regression" [11] by taking the ranking information into account. Rather than trying to classify the positive and negative examples, it shifts the focus in an attempt to maximize the gaps between each pair of positive and negative examples. Note that this is different from the margin maximization in the max-margin classifier which only considers the examples near the classification boundary. Formally, the model can be written as:

$$\max_{\lambda} \sum_{q \in Q} \sum_{d_1 \in D^+} \sum_{d_2 \in D^-} \log \sigma\left(\sum_{i=0}^{N} \lambda_i \left[f_i(d_1, q) - f_i(d_2, q) \right] \right)$$

where D_+ and D_- are the collections of positive/negative documents. It can be proven that the minimization of the disorder in the examples provides a lower bound of the average precision measure. However, optimizing the above loss function in a brute force manner could be computationally expensive in our case. For instance, the association between 100 positive and 900 negative examples results in an explosive 90,000 training pairs. Fortunately, we have come up with an approximation of the above loss function in form of:

$$\max_{\lambda} \sum_{q \in Q} \sum_{d \in D} w_d \log \sigma\left(\sum_{i=0}^{N} \lambda_i \left(f_i(d, q) - a_i \right) \right)$$

where w_d is the additional weights as ratio between the number of positive/negative data, and a_i is a shift factor automatically learned from the algorithms. It can be proved that this approximation is tight and the optimization complexity is the same as the standard logistic regression. Therefore, all of the following retrieval models will be built upon the approximated version of "ranking logistic regression."

2.3 Direct Query Expansion

The query-class based model has been demonstrated to be successful in several recent studies [1, 4, 10]. However, there is some query information that cannot be captured by the query-type representation. For example, the query "finding shots of tall buildings" has strong hints to suggest incorporating the output from semantic concept "buildings." But the limited number of query types cannot easily take this information into account. Therefore, we use an additional step to further refine the combination weights, i.e., when we find there is a direct match between the query description and the description of the semantic concepts, the corresponding concepts will be associated with a

positive weight, with a few examples from TRECVID 2005 shown in Table 1. In our current implementation, the additional concept weights are set to be equal to the weight of text retrieval. In our experiments, we only consider expanding the semantic concepts from the LSCOM-Lite concept list [7]. This keyword matching technique identified one additional concept each for 15 of the 24 TRECVID 2006 topics.

Table 1. Examples of TRECVID 2005 queries and corresponding expanded semantic concepts.

TRECVID 2005 Queries	Concepts
Find the *maps* of Baghdad	maps
Find one/more *cars* on the *road*	cars, roads
Find a *meeting* with a large table	meeting
Find one/more *ships* and *boats*	ship_and_boat

2.4 Empirical Results

The fully automatic search systems based on such techniques show steady improvement in TRECVID evaluations through the years, but fail to achieve the increased precision of a human user in the loop: mean average precision (MAP) for automatic systems lag well behind MAP of interactive searchers. Further, a topic-by-topic examination is necessary to better understand the current state of automatic search and its potential for helping an interactive user.

The MAP for **Automatic Baseline** was 0.045. The MAP for **Automatic** was 0.079. Stopping there, one would conclude that including the other modalities and visual concepts beyond text helps greatly, but a look at the average precision across topics in Figure 2 shows that most of the benefit from moving beyond the transcript text comes from the sports topic about soccer goalposts. For this topic, the average precision was 0.016 for **Automatic Baseline** but 0.552 for **Automatic**. Removing this topic from consideration drops down the difference between the two runs dramatically, with MAP for the remaining 23 topics now 0.046 and 0.058. Furthermore, a look at Figure 2 shows 4 specific topics naming people (Cheney, Saddam Hussein, Bush, C. Rice) where MAP for **Automatic Baseline** is 0.147 and MAP for **Automatic** is 0.153. Clearly, the text is contributing most of the necessary information to locate relevant shots for named person topics. Removing these topics, too, and the MAP drops precipitously close to zero, with MAP for the remaining generic non-sports topics being 0.025 for **Automatic Baseline** and 0.039 for **Automatic**. At this level of performance, is there any benefit to even considering the automatic runs for use in interactive systems?

Our hypothesis is yes, based on three arguments. First, Figure 2 shows that there are generic topics like snow or flames that possess characteristics where an automatic approach building from low level visual features and high level visual concepts can locate a reasonable set of relevant shots. Second, the evidence from MediaMill and Informedia work in the past show that an arsenal of tools – query-by-text, query-by-image, and query-by-concept – leads to better interactive performance than any single tool. Expanding the toolkit to include one more strategy, browsing the ranked shot output of the **Automatic** run which we label query-by-best-of-topic, is worth a look, since expanding the query arsenal before led to improvements. Third, the Carnegie Mellon XVR work in 2005 was motivated by a realization that the automatic search approaches may do poorly on average precision metrics grading the top 1000 shots per topic, but do reasonably well in recalling a number of correct shots in their top 4000 returns.

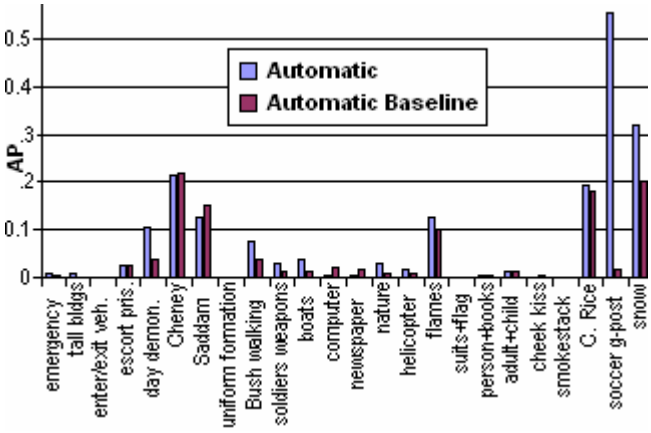

Figure 2. Average precision across 24 TRECVID 2006 topics by the two CMU automatic search runs.

Figure 3 shows a post-hoc analysis of the top 4096 returns from the **Automatic** run. If the order is left as is, but instead of 1000 we consider 2048, or 4096, the MAP actually grows only very slightly, but if a perfect "oracle" user would intervene and always shuffle relevant shots to the top of the graded shot set, then a perfect review and reordering of the top 4096 shots would result in a MAP of 0.510. There is adequate recall in the automatic search runs, but with depths greater than 1000 shots. The role of the experiments here with **Informedia Full, Informedia Limited, XVR-1,** and **XVR-2** was to see whether the inclusion of additional query tools mattered, whether the interface mattered, and based on how well these systems compared to other TRECVID 2006 interactive systems, whether indeed there was a benefit in passing the automatic run results onto the user for interactive review and filtering.

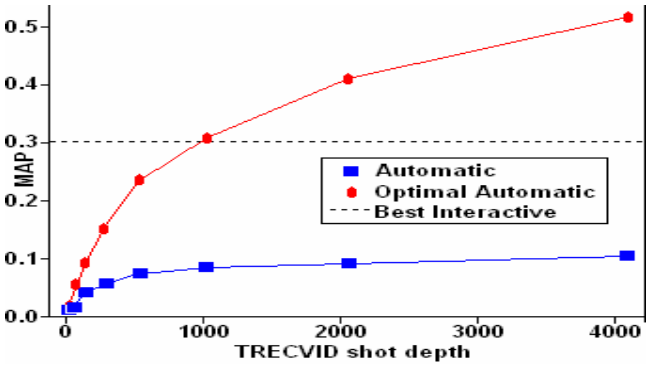

Figure 3. MAP for Automatic run and a perfect reordering of that run (putting relevant shots first), at different shot depths.

3. FROM AUTOMATIC TO INTERACTIVE

The **Informedia Full** treatment allows users to retrieve and view shots based on all 4 strategies, as reflected in Figure 4. Aside from this addition of query-by-best-of-topic, the interface is the same Informedia storyboard one used in TRECVID 2005 and reported at CIVR 2006 [3]. Figure 5 emphasizes the new addition.

The **Informedia Limited** run has the same interface as shown in Figure 5, with a video review option (video window shown in lower left) and the captured shots collected and shown in the partially cropped right pane. The big difference is that we wanted to test the utility of the **Automatic** run output, and so the **Informedia Limited** run only could access shots through "query-by-best-of-topic", with all other query functionalities (query-by-text, by-image, by-concept) inaccessible. Likewise, the other 2 interactive runs have access only to the output of the automatic search run, except that they use extreme retrieval (XVR) interfaces rather than the traditional Informedia dense storyboard displays. Figure 4 shows the interactions breakdown for the different treatment groups following an analysis of the transaction logs for the accumulation of 15-minute topic sessions. Figure 4 is interesting in that the dominance of the query-by-text strategy for the **Informedia Full** interface is much less than in prior years [2, 3], with the rich query functionality allowing for a diversity in interactions: text query accounts for only 16% of captured shots.

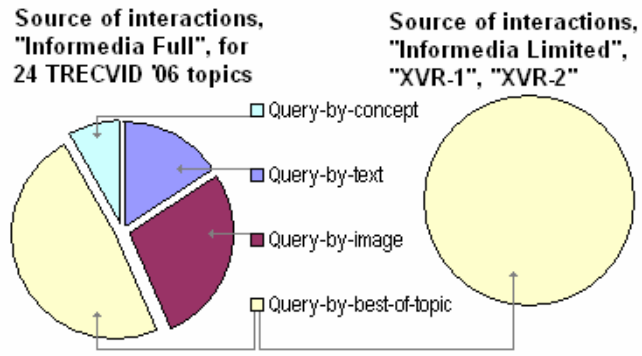

Figure 4. "Informedia Full" use of 4 query strategies.

Regarding the "query-by-concept" functionality, we worked only with the LSCOM-Lite concepts [7]. If we had a richer palette of LSCOM concepts, e.g., if rather than just 39 LSCOM-Lite we had implemented access to 834 in the full LSCOM set [7], then perhaps query-by-concept functionality would have received a greater percentage of attention in the **Informedia Full** treatment, and perhaps also led to better ranked shot sets for the query-by-best-of-topic group, since the automatic run utilizes semantic concepts as well but for TRECVID 2006 worked only with LSCOM-Lite. This is a point for future work, but in the experiments reported here we wished to see specifically whether bringing in query-by-best-of-topic would produce benefits, without also changing the nature of query-by-concept from how we used it with TRECVID 2005.

The **Informedia Full** run, **Informedia Limited** run, and **XVR-1** run scored as 3 of the top 5 TRECVID 2006 search runs as rated by mean average precision (MAP) across the 24 search topics. From these successful performances, we conclude that the ranked shot lists from automatic search are quite useful as starting points for interactive review and filtering. The benefits of additional interactive query capability are shown by the improved performance of **Informedia Full** treatment over the other treatments, with it scoring significantly better than all other TRECVID search runs.

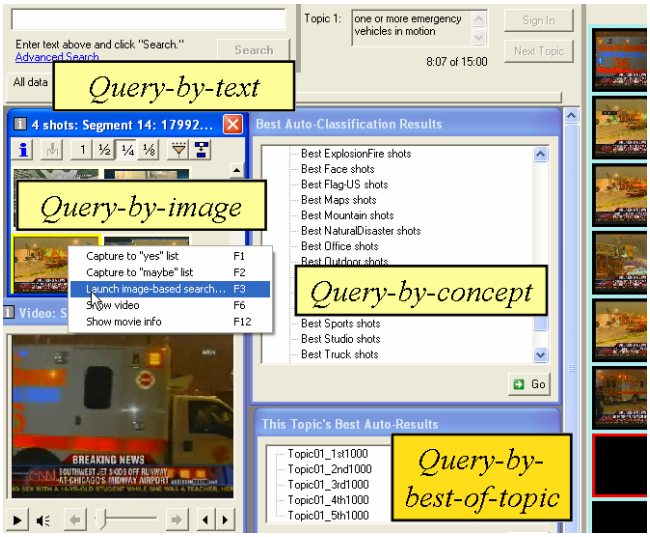

Figure 5. Screen from "Informedia Full" interface, with 4 query options to produce shots sets for viewing, a video preview window to play back video sequences (lower left), the query description and topic timer (upper area), and capture panel (right) for answers.

In addition, the XVR interfaces are noted as requiring the absolute attention and focus of the user for optimal performance [5] and hence being more stressful and "extreme." The traditional Informedia storyboard interface has received consistent high scores for end-user satisfaction [2, 3]. The less stressful Informedia storyboard runs produced better performances than the XVR runs on the TRECVID 2006 search topics. Figure 6 shows the mean average precision for the top 44 scoring runs using MAP as the metric, with 79 other TRECVID 2006 runs scoring below our **Automatic Baseline** run. The highest scoring automatic run achieved a MAP of 0.087, again confirming that interactive searchers can produce better performance than a completely automatic approach.

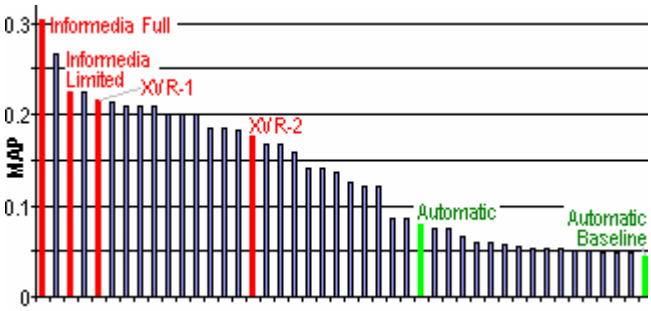

Figure 6. MAP for best 45 TRECVID 2006 search runs.

4. INTERFACE DETAILS & DISCUSSION

The TRECVID test corpus consisted of 165 hours of U.S., Arabic, and Chinese news sources. This collection was automatically divided into shots at the Fraunhofer (Heinrich Hertz) Institute in Berlin, producing a common shot reference of 79,484 TRECVID shots used in scoring across TRECVID 2006 participants [6]. We expose a richer visual set of 146,328 Informedia shots in our interface by representing both the shots and "sub-shots" in the common shot reference: Informedia shots never cross common shot reference boundaries, but some TRECVID shots are represented by more than one Informedia shot. We account for visual diversity within a TRECVID shot at the expense of having more overall shots (146,328) to manage in the interface.

The **Informedia Limited** and **Informedia Full** interfaces both made use of the following strategies to maximize the number of Informedia shots reviewed by the human user. First, a dense packing of shot thumbnails in storyboards was the primary interface. The shots are always clustered by story segment to better preserve the temporal flow within news stories and facilitate an easier left-to-right, top-to-bottom scan of the thumbnails. Five storyboards were available for query-by-best-of-topic, holding the top-scoring set of 1000 Informedia shots from the automatic search run, the next 1000, and so on. Within each set of 1000, shot reordering was done to cluster by story segment. Users had control over the thumbnail resolution, row and column count for the storyboard, and could impose additional filtering on the storyboard, e.g., to show just thumbnails with faces or any of the other LSCOM-Lite concepts, where the threshold is set by the user through a dynamic query slider [2]. Figure 7 shows an example use of such sliders to reduce a set of shots from a text query on "emergency ambulance police" down to those shots automatically tagged with some confidence as being cars and road shots.

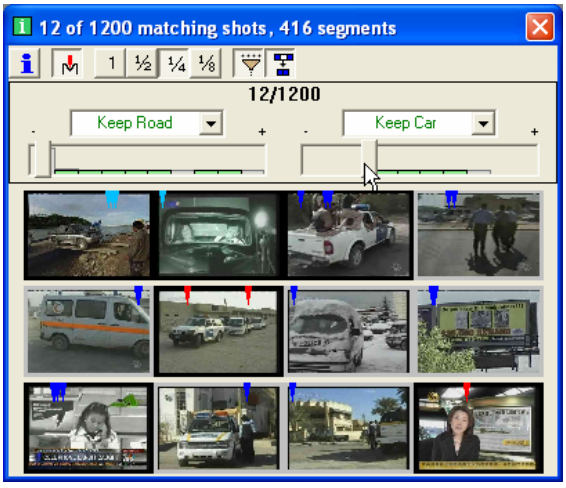

Figure 7. Storyboard with dynamic query sliders to subset the shown thumbnails to those automatically tagged as being car shots and roads shots, here aggressively reducing 1200 shots down to 12. Colored notches indicate temporal locations of color-coded matching text from query.

The user's interaction history was tracked so that once a shot was "captured" into the answer set shown at the right pane of Figure 5, it would not be shown again in subsequent storyboards. Likewise, if a shot was "overlooked" and skipped over while a later shot in the storyboard was captured, the overlooked shot would not be listed again in subsequent storyboards. This implicit tagging of shots as "irrelevant, so never show again" was found to be too aggressive in TRECVID 2005 evaluations [3]. For TRECVID 2006, we kept its use in the interface the same as in 2005, to filter what the user sees, but relaxed its use in automatic expansion from the user's captured set of N shots out to the submitted set of 1000 shots for the topic at hand. In TRECVID 2006, we did not consider and filter by the overlooked shot set during automatic expansion out to 1000.

Navigation mechanisms let the user select any thumbnail and capture that shot, show the video queued to that shot's start time, show the storyboard of thumbnails for other sibling shots in the same story, or show the story and broadcast information in a text tree view. Informedia story segmentation was used to cluster shots into segments, with segments being the unit of information retrieval for the query-by-text option available in **Informedia Full**. Also in **Informedia Full** only, the user could launch a color-based image search, query-by-image, from any thumbnail. The consistency and ease of access for these shot-based operations was found to have a direct impact on usability in prior studies [2, 3]. The new work here was the investigation into the utility of query-by-best-of-topic.

Two users contributed to the **Informedia Limited** run, and one user to the **Informedia Full** run. These three users were experts in the Informedia interface, contributing to its development and use, but were isolated from the TRECVID 2006 topics and data set.

The Informedia storyboard interfaces allows for impressive numbers of shots to be reviewed by motivated users within the 15-minute time limit. Figure 8 shows a conservative count of the number of Informedia shots either marked as correct, or passed over when marking another shot as correct, during the 15-minute topic run. Figure 9 shows the AP for the topics and interactive runs. The review count shown in Figure 8 is conservative in that if the user finishes off with a review of a storyboard with, say, 78 irrelevant shots at the end of the storyboard that get no user action, those 78 are not counted because the user neither marked any as correct nor passed over them and marked a following one as correct. Of course, with better transaction logging we could get more accurate counts, but even these lower bound numbers are impressive in that the interface supports review of thousands of shots within the time limit. Clearly, some topics take more time to review than others, e.g., both "emergency vehicles in motion" and "a kiss on the cheek" were time-consuming. The average number of Informedia shots reviewed per topic with XVR-1 (manual browsing with resizable pages) was 1314, and with XVR-2 (RSVP) was 1364.

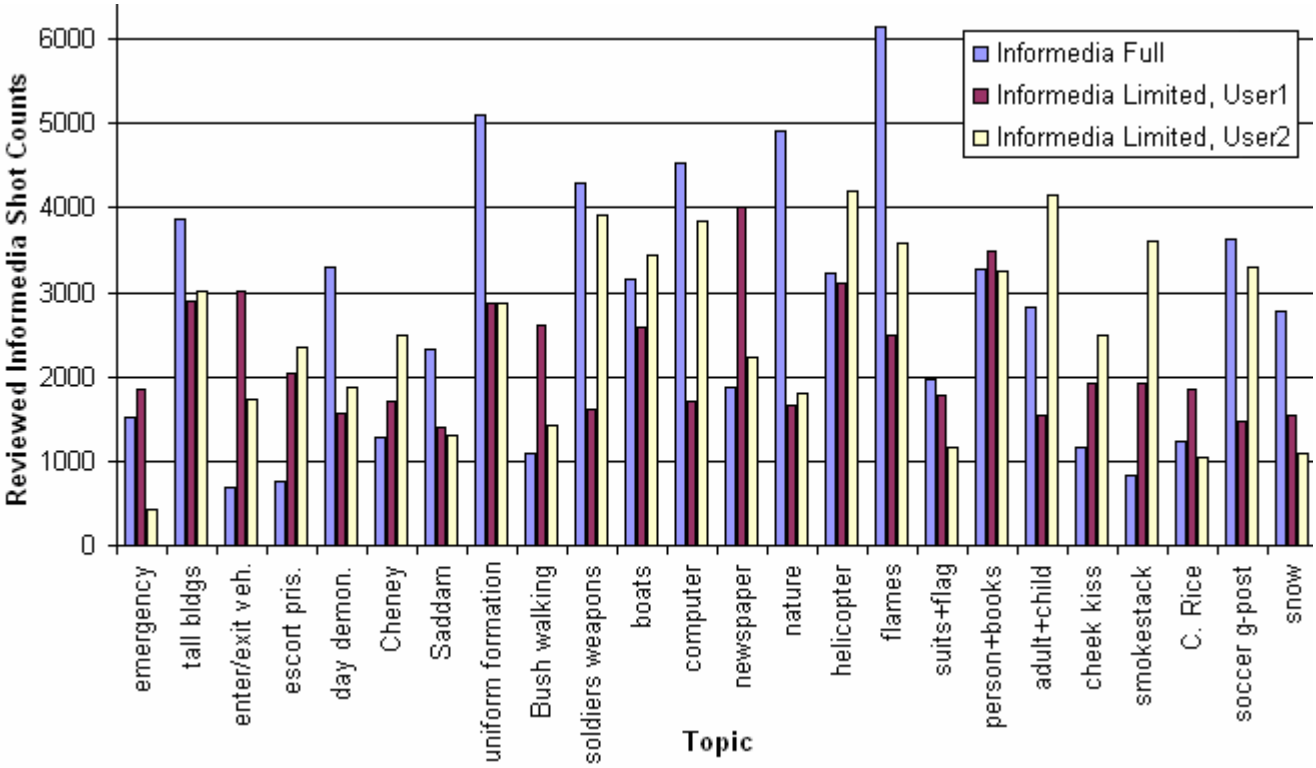

Figure 8. Count of Informedia shots reviewed in 15 minutes per TRECVID 2006 topic (2 different users with "Informedia Limited", with a third user interacting with "Informedia Full").

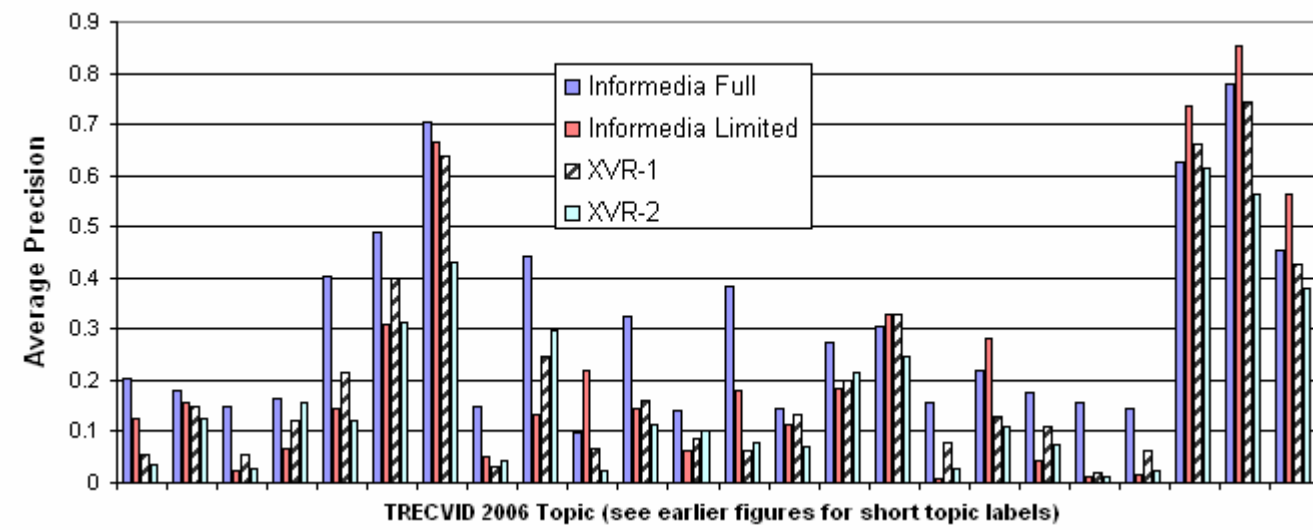

Figure 9. Average Precision across TRECVID 2006 Topics for 4 CMU Interactive Search Runs.

Using the described storyboard interfaces, one user with Informedia Limited reviewed 2195 shots per topic, the other 2526. Finally, the Informedia Full user reviewed 2740 shots per topic. By providing an interface enabling thousands of shots to be visually inspected in an efficient manner, the user is armed with a system allowing for human correction of "poor precision but adequate recall" shot lists. The automatic search provides such shot lists, resulting in an overall benefit for providing the query-by-best-of-topic capability into the Informedia storyboard interface, as evidenced by the results shown in Figure 6.

5. CONCLUSION

In prior TRECVIDs, the fully automatic search run succeeded with good recall for many of the topics with no user in the loop, but the relevant shots were distributed throughout the top 3000 to 5000 slots in the ranked shot list, causing the average precision (AP) for the automatic search run to lag well behind the AP scores for the best interactive runs. By relying on an intelligent human user possessing excellent visual perception skills to compensate for comparatively low precision in automatically classifying the

visual contents of video, a human user can filter the automatic set and produce a set that retains most or all of the relevant shots from the automatic set, but with much greater precision. We compared three different interfaces to see which one maximizes the human efficiency in labeling relevant shots from the ranked list. The traditional Informedia storyboard interfaces outperform the extreme video retrieval (XVR) interfaces, with rapid serial visual presentation lagging significantly. These results confirm our suspicion that the success enjoyed in TRECVID 2005 by the XVR interfaces was due not to the interface style but rather to the use of good seed information in the form of ranked shot lists from automatic search.

Furthermore, the Informedia traditional style of storyboard display allows for a greater number of shots to be visually reviewed by a human user within the 15-minute TRECVID time limit, without nearly as much stress as the XVR interfaces. When additional query capabilities are provided in the interface, performance improves significantly, as evidenced by the top score achieved by the **Informedia Full** run shown in Figure 6.

The traditional Informedia "let the user control all" storyboard-based strategy may benefit from incorporating some of the XVR "system controls all" ideas to increase the reviewed shot count a bit, e.g., starting off topics with XVR automated scrolling but letting the user interrupt and take over at any point. Reviewing more shots lets the human user compensate for shot-ordering solutions that are suboptimal but still have adequate recall. Looking back on Figure 3, though, indicates that pushing the number of shots reviewed from 2740 to 4000 may only marginally increase overall MAP. Instead of focusing on getting even more shots to the users' eyes, a more promising approach for news shot-based retrieval is to enhance the query arsenal, in particular query-by-concept.

The utility of query-by-concept may increase substantially as the concept set grows from the LSCOM-Lite set of 39 to the LSCOM set approaching 1000 concepts. There will likely be a need for automated support to help the user decide in ordering query strategies once the concept space broadens to this level, e.g., which concepts to examine first, in what combination, given a query. The impact on relevance feedback to reweight concepts and change shot ordering, and use of machine learning to thin concept options to a smaller recommended set for a given topic are worth further examination once the full LSCOM concept set is employed.

TRECVID is extremely valuable as an open, metrics-based evaluation forum for video retrieval. Prior studies on the interactive search task by TRECVID participants have shown the value of enhancing the interactive arsenal with query-by-text, query-by-image, and query-by-concept approaches. The work reported here shows the additional positive impact provided by adding in a query-by-best-of-topic option as well.

6. ACKNOWLEDGMENTS

This work is supported by the National Science Foundation under Grant No. IIS-0205219. The search runs detailed here would not be possible without the efforts of Robert Baron and Bryan Maher. Details about Informedia research and the full project team can be found at www.informedia.cs.cmu.edu. Our thanks to NIST and the TRECVID organizers for enabling this video retrieval evaluation work.

7. REFERENCES

[1] Campbell, M., Ebadollahi, S., Naphade, M., Natsev, A., Smith, J.R., Tesic, J., Xie, L., Haubold, A. IBM Research TRECVID-2006 Video Retrieval System. In *NIST TREC Video Retrieval Online Proceedings.* NIST, 2006, //www-nlpir.nist.gov/projects/tvpubs/tv6.papers/ibm.pdf.

[2] Christel, M., and Moraveji, N. Finding the Right Shots: Assessing Usability and Performance of a Digital Video Library Interface. In *Proc. ACM Multimedia* (New York, NY, Oct. 2004), ACM Press, New York, 2004, 732-739.

[3] Christel, M., and Conescu, R. Mining Novice User Activity with TRECVID Interactive Retrieval Tasks. In *LNCS 4071: Proc. Image and Video Retrieval (CIVR)* (Tempe, AZ, July 2006), Springer, Berlin, 2006, 21-30.

[4] Chua, T.-S., Neo, S.-Y., Li, K.-Y., Wang, G., Shi, R., Zhao, M., Xu, H. TRECVID 2004 Search and Feature Extraction Task by NUS PRIS. In *NIST TREC Video Retrieval Online Proceedings.* NIST, Gaithersburg, MD, 2004, http://www-nlpir.nist.gov/projects/tvpubs/tvpapers04/nus.pdf.

[5] Hauptmann, A.G., Lin, W.-H., Yan, R., Yang, J., and Chen, M.-Y. Extreme video retrieval: joint maximization of human and computer performance. In *Proc. ACM Multimedia* (Santa Barbara, CA, Oct. 2006), ACM Press, New York, NY, 2006, 385-394.

[6] Kraaij, W., Over, P., Ianeva, T., and Smeaton, A. TRECVID 2006 - An Introduction. In *TRECVID Online Proceedings*, //www-nlpir.nist.gov/projects/tvpubs/tv6.papers/tv6intro.pdf.

[7] 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.

[8] Snoek, C., Worring, M., Koelma, D., and Smeulders, A. Learned Lexicon-Driven Interactive Video Retrieval. In *LNCS 4071: Proc. Image and Video Retrieval (CIVR)* (Tempe, AZ, July 2006), Springer, Berlin, 2006, 11-20.

[9] Snoek, C., Worring, M., Koelma, D., and Smeulders, A. A Learned Lexicon-Driven Paradigm for Interactive Video Retrieval. *IEEE Trans. Multimedia 9*(2), Feb. 2007, 280-292.

[10] Yan, R., Yang, J., and Hauptmann, A. Learning Query-Class Dependent Weights in Automatic Video Retrieval. In *Proc. ACM Multimedia* (New York, NY, Oct. 2004), ACM Press, New York, NY, 2004, 548-555.

[11] Yan, R., and Hauptmann, A.G. Efficient margin-based rank learning algorithms for information retrieval. In *LNCS 4071: Proc. Image and Video Retrieval (CIVR)* (Tempe, AZ, July 2006), Springer, Berlin, 2006, 113-122.