9-2007

Exploring Concept Selection Strategies for Interactive Video Search

Michael G. Christel
Carnegie Mellon University

Alexander Hauptmann
Carnegie Mellon University

Published In
Semantic Computing, 2007. ICSC 2007. International Conference on , 344-354.

Exploring Concept Selection Strategies for Interactive Video Search

Michael G. Christel and Alexander G. Hauptmann
School of Computer Science, Carnegie Mellon University, Pittsburgh PA USA
{christel, hauptmann}@cs.cmu.edu

Abstract

Ranked shot lists from 39 automated LSCOM-Lite concept classifiers are investigated with respect to 24 TRECVID 2006 topics. Selecting the best fitting concept or pair of concepts produces the shot set with greatest utility, rather than drawing fewer shots from a larger set of concepts. Mean average precision measures show concept-based shot sets have great utility for topics when perfectly traversed by a user. Using empirical data, however, shows that realistic ability to separate relevant shots from irrelevant ones and recall all the relevant ones is topic-dependent and far from perfect. Concept-based strategies including user-driven selection strategies not using idealized oracle prioritization are also discussed, with implications for query-by-concept in interactive video retrieval as concept spaces grow from tens to thousands.

1. Introduction

NIST TRECVID video retrieval experiments through the past four years have benchmarked the progress of interactive search against broadcast news corpora [1]. These experiments have shown that systems providing text search against transcripts of the spoken narrative, coupled with additional query capabilities for low level image attributes and higher level semantic concepts, score significantly better than other interactive systems [2, 3, 4, 5, 6]. In particular, the University of Amsterdam MediaMill team [4, 5] and Carnegie Mellon University Informedia team [3, 6] have repeatedly demonstrated the success of systems providing three common means of querying the news corpus to produce shot sets: query-by-textual-keyword, query-by-image-example, and query-by-concept.

Prior TRECVID work has established that a human searcher in the loop significantly outperforms fully automated news video search systems without such a human searcher [1]. Prior work has also established the importance to this point of promoting access to all three query capabilities (text, image, concept) for more effective search and higher mean average precision across the TRECVID topics [7]. However, there may be instances or corpora for which a particular query capability is severely handicapped or impossible. For example, if the corpus is a foreign news source for which there is no provided closed-caption transcript and no automated speech recognition engine exists to provide translation into text, then there will be little or no text to search. If the video feed is only visual with no audio channel there will be no spoken narrative to translate into text. In such cases, the other query forms will need to compensate for the missing query-by-text functionality.

Image-based queries by color, shape, and other low level image syntactical attributes has been studied extensively in the content-based image retrieval community and remains difficult due to the semantic gap produced by "the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation" [8]. Systems like QBIC retrieve images based on attributes like color and texture [8], but studies have questioned the utility of image searching according to such low-level properties [9]. The semantic gap makes it difficult for the user to formulate queries against imagery and video. CBIR methods that rely on higher-level semantic features, perhaps organized into a video ontology [10, 11] that is sensible for a user community, can improve user understanding and bridge the semantic gap. An ontology – a powerful way to describe objects and their relationships to other objects – can be better than keywords for retrieval from a video library, because a general information need can be satisfied by the ontology even without exact matches to provided keywords [12]. A concern is that automated classifiers for such higher-level features are much less accurate than those that detect color, shape, or texture low-level features. An open question is the effect of feature

availability, applicability, and accuracy of automated feature classification on video retrieval [13], which forms part of the motivation for our work.

This paper makes use of TRECVID 2006 data, specifically the 24 TRECVID 2006 topics defined by NIST and the NIST-provided pooled truth for these 24 information needs, i.e., topics [1]. The pooled truth allows shot sets to be evaluated with respect to a given topic. The question of focus in this paper is as follows: **If users only had access to query-by-concept, i.e., access to the ranked shot sets from 39 fully automatic concept classifiers, how well could they do on the topics?**

Our investigation builds from the Large Scale Concept Ontology for Multimedia (LSCOM) [11]. We make use of the Carnegie Mellon University classifiers for the 39 LSCOM-Lite concepts listed in TRECVID 2006's call for participation. The accuracy of the classifiers was assessed for a subset of 20 concepts. These 20 evaluated concepts included sports, weather, office, meeting, desert, mountain, waterscape, corporate leader, police/security forces, military personnel, animal, computer or TV screen, US flag, airplane, car, truck, people marching, explosion/fire, maps, and charts, which were found at medium frequency in the training data. The best detection system for the semantic concepts that were evaluated obtained mean average precision (MAP) of about 0.19 over the 20 concepts. However, other systems performed better on individual concepts. The CMU system was evaluated at a performance of around 0.16 MAP just behind systems submitted by Tsinghua University [14] and IBM [15], but above the rest of the 30 participants in the TRECVID 2006 evaluation.

In this paper, we do not investigate the effects of improving accuracy beyond these currently achievable levels. Rather, we look into the question of whether the user, if provided with the ideal automatic recommendation system that points out which concept(s) to use for which topic, would be able to perform sufficiently well with only an inspection of the top-ranked shots for the recommended concept or concepts. Section 2 discusses the framework of the experiment. Section 3 presents results in an ideal interaction framework. Section 4 revisits those results in a more realistic framework informed by empirically collected data with respect to the TRECVID topics. Section 5 presents conclusions and future work.

2. TRECVID 2006 search, employing only the concepts

Suppose the user only had access to the 39 LSCOM-Lite concepts. What would they do with them in the TRECVID interactive search framework? In this framework, they have 15 minutes with which to review the automatically produced shot sets for the different concepts. One thing they might do is look at the corpus as a whole, all 79484 shots, and then start filtering down the shots to just those having certain concepts like "road" and not other concepts like "face." The dynamic filtering of shot sets based on an intersection of a small set of just 6 concepts was found to be quite complex, with concept utility in filtering greatly diminished as the accuracy of the concept classifiers drops [16]. Users rarely made use of concepts as a filtering strategy in Carnegie Mellon TRECVID search runs [2, 3, 6, 7, 17] because it sacrificed recall. Instead, in these prior experiments involving over 60 users, the participants opted for a broader inspection strategy, looking through thousands of shots for many topics. We leave the use of concepts as intersecting filters, including the use of concepts for exclusion rather than inclusion (e.g., show all shots NOT tagged as roads) for future experiments. Here, we consider the use case where the human searcher can inspect one or more concepts given a topic, as explained below.

For TRECVID 2006 we ran three users through all 24 topics and collected transaction logs confirming the broad inspection strategy seen in prior TRECVID interactive runs validated with tens of users [3, 6, 7, 17]. These three users inspected shots represented by thumbnail imagery in storyboard layouts and made their judgments almost exclusively from such thumbnails [6], because the time penalty for carefully reviewing the video associated with each shot was too great. Such inspection allowed 2740, 2526, and 2195 shots to be reviewed on average, per topic, in the 15-minute time limit by these three users. From this data, we conservatively set an "anticipated reviewed shot count per topic" constant value of 2000 for use in this investigation. We expect that users will be able to review 2000 shots and judge them for relevance to a given topic in the 15-minute time limit.

The remaining question is what 2000 shots to show: the top-ranked 2000 from one concept, or top 1000 from two different concepts, or the best 125 from 16 concepts? What concepts should be selected? To investigate the first question, we make use of perfect wisdom and wise oracle selection in answering what concepts to select: pick the concept or concepts that are best suited to the topic at hand, based on pooled truth data.

Our procedure is as follows. Consider the highest ranked 2000 shots for each of the 39 concepts by the automatic classification, using the empirical-based constant of 2000. For each topic, grade the top 2000

on a recall-at-2000 (R2000) metric against pooled truth, i.e., how many of the top 2000 shots are actually correct relevant shots for the topic. Also collect recall-at-125 (R125), R500, and R1000 for use in generating other shot sets.

When selecting a single concept for a topic, pick the one having the highest R2000 value. When selecting two concepts, pick the two with highest R1000 for a topic. Then, take the first 1000 shots for the highest scoring R1000 concept. Take the rest of the 1000 shots from the second-highest scoring R1000 concept, perhaps going deeper than the top 1000 shots for that concept because some may already be represented in the candidate set from the first-used concept. The result is a 2000 shot set produced from the top 2 concepts based on R1000 (graded with NIST pooled truth for each topic). Similarly, generate candidate shot sets of 2000 shots from 4 concepts based on R500, 8 concepts based on R250, and 16 concepts based on R125. Finally, produce a set of 2000 shots drawing the top-ranked 51 or 52 shots from each of the 39 LSCOM-Lite concepts.

Note that this procedure minimizes the effects of concept ordering. For example, for a topic best served by concepts "Explosion-Fire" and "Road", the top 1000 explosion shots are gathered, followed by the top 1000 road shots not already represented in the kept 1000 explosion shots. This set of 2000 shots will be nearly identical in membership to a reordering that would keep the top 1000 shots from "Road", followed by the top 1000 shots from "Explosion-Fire" not already represented in the kept 1000 road shots. Since we are not concerned about shot ordering within the set of 2000, assuming equal inspection of all shots in the set limited to 2000 items, the ordering of concepts is also less critical. Table 1 lists the concepts used for each of the topics based on this strategy and the CMU concept classifiers, for completeness noting the concept ordering as well (2a first, then 2b; 4a followed by 4b, 4c, 4d). Reordering the concepts (2b, then 2a; or 4d-c-b-a; similar reverse order for 8 and 16 concepts) produces 91.6%, 92.0%, 92.2%, and 91.8% of the same shots as the listed orderings, respectively.

Table 1. Through perfect oracle assignment, the "1" most relevant concept per topic as judged by recall at depth 2000 (R2000); most relevant 2 concepts "2a" and "2b" based on R1000; most relevant 4 via R500.

	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196
1	13	9	30	24	35	28	28	24	28	24	17	31	27	11	36	36	6	28	18	21	33	25	1	15
2a	36	9	13	24	35	28	28	24	28	24	17	31	27	11	29	36	6	28	18	21	33	25	1	15
2b	13	16	30	23	18	21	4	35	34	8	33	5	7	12	36	37	21	31	34	4	17	28	11	37
4a	13	9	30	24	35	28	28	24	28	24	17	31	27	11	29	36	6	28	34	4	33	25	11	15
4b	30	16	13	35	18	21	4	23	34	8	33	5	7	12	36	37	31	33	18	21	17	18	1	37
4c	36	14	14	23	24	25	34	8	21	36	12	15	11	17	10	29	21	15	35	25	12	28	34	29
4d	24	13	23	25	8	35	25	35	18	10	14	20	16	37	33	8	18	9	11	20	36	34	38	8

Table 2 and Table 3 list the 39 concepts and 24 topics. Clearly, there are some expected fits, such as C36 "explosion-fire" and C13 "road" for the top two concepts for Topic 173 on emergency vehicles, as well as some unexpected fits such as C31 "bus" for the Topic 184 computer screen query. The latter anomalies are due to a small concept corpus of only 39 which may not contain any or many good fits for a topic, and the low accuracy of some automated concept classifiers.

With the data of Table 1 we are prepared to run our investigation. If the user just looked at the 2000 shots returned by the following strategies – 1-concept, 2-concepts, 4-concepts, 8, 16, and all 39 – there would be differing numbers of correct shots within the presented 2000 shots. These counts are presented in Table 4.

Table 2. 39 LSCOM-Lite concepts.

1	Sports	20	Person
2	Entertainment	21	Government-Leader
3	Weather	22	Corporate-Leader
4	Court	23	Police-Security
5	Office	24	Military
6	Meeting	25	Prisoner
7	Studio	26	Animal
8	Outdoor	27	Computer-TV
9	Building	28	Flag-US
10	Desert	29	Airplane
11	Vegetation	30	Car
12	Mountain	31	Bus
13	Road	32	Truck
14	Sky	33	Boat-Ship
15	Snow	34	Walking-Running
16	Urban	35	People-Marching
17	Waterfront	36	Explosion-Fire
18	Crowd	37	Natural-Disaster
19	Face	38	Maps 39 Charts

Table 3. 24 TRECVID 2006 topics.

173	emergency	185	newspaper
174	tall buildings	186	nature
175	enter/exit vehicle	187	helicopter
176	escort prisoner	188	flames
177	day demonstration	189	suits+flag
178	Cheney	190	person+books
179	Saddam	191	adult+child
180	uniform formation	192	cheek kiss
181	Bush walking	193	smokestack
182	soldiers weapons	194	C. Rice
183	boats	195	soccer g-post
184	computer	196	Snow

3. Results with ideal user interaction

A closer inspection of Table 1 shows that once a concept is found to be a good fit for a topic, it tends to stay a good fit, even if we only consider recall at depth 125 or 500 or 1000. Occasionally, though, concept membership does change. For example, with the helicopter topic, Topic 187, the concept returning the most helicopter shots in the top 2000 is C36 "explosion-fire", but the concept with the most helicopter shots in its top 1000 shots is C29 "airplane" followed by C36. Similarly, the best topics for the uniform formation topic, Topic 180, do not include C23 "police-security" when only taking the best one or two concepts, but when taking the best four looking at R500, C23 returns the second-most relevant shots in its automatically ranked top 500.

The interactive search runs with human user involvement have significantly outperformed fully automated TRECVID search runs because the human viewer is very good at quickly judging through inspection the relevance of an image to the topic at hand. Suppose at first that the user is perfect in interactive review, i.e., that after looking through the 2000 shots they mark all of the relevant ones for the topic (perfect recall), and only mark the relevant ones (perfect precision). For example, for Topic 173 using the 2000 Road shots they find and mark all 31 relevant shots and only those shots. For the same topic and the 2000 shots derived from Explosion-Fire and then Road they find and mark all 37 relevant shots. Such an exercise leads to the average precision numbers reported in Figure 1 for the different concept combinations 1, 2, 4, 8, and 16.

Clearly, these represent ideal cases for two reasons. First, the best concept or set of concepts are applied to each topic through oracle inspection of the truth data. Via pooled truth from NIST for TRECVID 2006, we establish the ideal concepts per topic given the CMU classifiers, e.g., using the CMU-classified shot sets for "military" and "outdoor" for the "soldiers weapons" topic. Second, the average precision is computed assuming perfect precision and recall by the user: the user will pick out only the relevant shots for a topic from the set of 2000, and be able to inspect all 2000. For some topics, like soccer, this is reasonable, because most people can quickly and accurately select the soccer thumbnails from sets of sports thumbnails. For other topics, however, like entering and exiting a vehicle, simple inspection of thumbnail imagery in storyboard layouts (the interface design allowing for 2000 shots to be inspected within the TRECVID topic time limit) will not be suitable for such high precision and recall.

Table 4. Counts of correct, relevant shots for a topic across 24 TRECVID 2006 topics and 2000-shot set generated via 1 concept, 2 concepts (1000 shots each), 4 (500), 8 (250), 16 (125), and all 39 (51 or 52 shots each); also the counts from the three strategies not requiring an oracle ideal concept selection: BF1 (breadth-first until one found), RW (recent window) or Y (overall yield) strategy

	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196
1	31	237	41	22	165	26	37	51	25	140	112	44	66	113	42	90	75	16	80	15	18	16	262	211
2	37	206	42	29	149	21	28	49	23	126	98	49	47	139	44	73	50	19	66	10	15	18	269	194
4	31	193	43	27	118	21	16	51	18	128	92	40	28	134	49	52	41	12	52	7	13	13	237	178
8	28	151	36	17	86	14	9	46	16	108	83	25	21	116	44	45	34	8	46	4	13	8	193	151
16	18	117	28	15	54	11	2	34	9	75	67	16	13	88	31	35	16	5	26	2	9	1	145	129
39	9	60	15	9	33	6	1	19	5	41	33	10	3	54	18	19	6	2	10	1	4	1	85	67
BF1	9	234	21	12	148	11	0	38	12	125	72	18	17	94	36	72	25	2	11	1	5	1	251	185
RW	11	210	17	17	138	26	1	29	23	91	63	26	22	76	29	19	45	4	6	4	17	1	227	83
Y	7	235	19	19	157	22	0	46	20	137	68	23	37	90	31	85	50	1	11	1	12	1	264	209

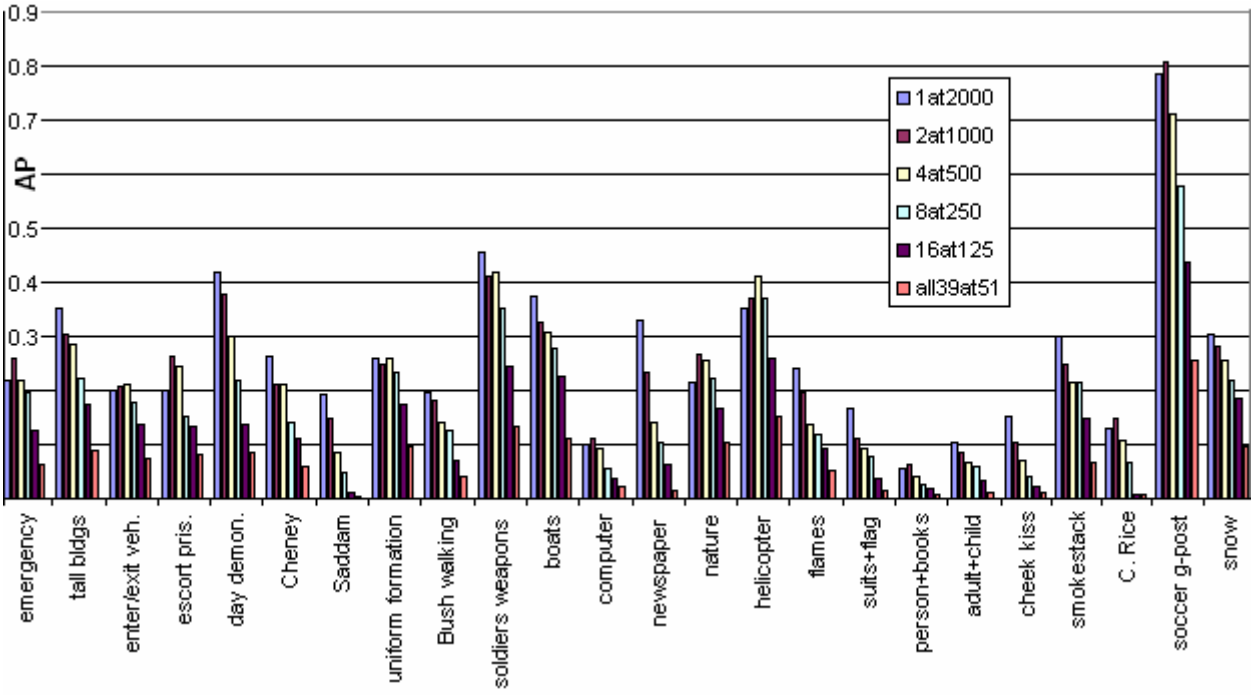

Figure 1. Average Precision (AP) across 24 TRECVID 2006 topics assuming perfect tagging of the correct shots (counts in Table 4) from the candidate set of 2000 generated from strategies listed in Table 4.

Three other strategies was also explored, which removed the assumption that an oracle was available to help select the most appropriate concept. Instead, the hypothetical user adopted either: (**BF1**) – a strategy where all topics were explored in a breadth first manner until a relevant shot was seen in one of the concept rankings, and from then on only this concept was used; (**RW**) – an approach which counted the relevant shots found in a window of the most recent shots in the rankings for concept and explored the concept with the highest recent window count; or (**Y**) – a yield approach which examined the ratio of relevant shots to all shots examined for a concept shot list, and choosing the highest yielding concept to explore next. In each of these methods, the user would initially look at all concepts equally, until the approach selected the most "promising" concept list to explore next.

Table 5 summarizes the results through mean average precision (MAP), bringing in 3 other measures as well for comparison: the fully automated search run and the top-scoring TRECVID 2006 interactive search run by Carnegie Mellon, and the average of the top 10 (interactive) graded search runs for TRECVID 2006.

We did check that the value of the reported approach is not due solely to oracle selection of the best shot set from 39 candidates. We randomly produced 39 sets of 2000 shots each, and scored them and ordered them as outlined previously in Section 2. The MAP for drawing the best "random concept" per topic is a miniscule 0.050, well below choosing the best LSCOM-Lite concept and its MAP of 0.265. Hence, with 39 sets, the power is coming from the concept classifier rankings. In future work, when considering the full concept ontology and LSCOM set sizes of 800 or more concepts rather than 39 [11], such an additional check against 800 plus random shot sets, one per concept, is warranted. It ensures that oracle selection analyses are not biased by sampling anomalies brought on by selecting from hundreds of candidate shot sets.

A few points are immediately obvious from Tables 4 and 5 and Figure 1. First, looking only at the concatenation of the top-ranked shots from all 39 concepts and using that same list of 2000 shots regardless of topic is not a very good strategy. Its MAP of 0.069 is worse than the fully automated run scoring an MAP of 0.079, and pales significantly compared to the top-10 graded runs' MAP of 0.228. This is not surprising, in that many of the concepts listed in Table 2 make little sense together, e.g., there will rarely be any topic requiring both "studio" and "natural disaster." Some concepts like "studio" are almost universally an exclusionary concept: a topic's relevant shots will never be studio shots. In this investigation, we simplified concept combinations and so only look at including shots in a set based on

automated concept rankings, disregarding the power of exclusion. It is clear that the all39 strategy fails, lacking any tuning to individual topics.

Table 5. MAP across various TRECVID 2006 runs (first 9 rows, also in Fig. 1, based on 100% user accuracy in identifying and keeping relevant shots).

Run	MAP
1at2000, 100% user accuracy	0.265
2at1000, 100% user accuracy	0.248
4at500, 100% user accuracy	0.220
8at250, 100% user accuracy	0.179
16at125, 100% user accuracy	0.127
all39at51, 100% user accuracy	0.069
BF1, 100% user accuracy	0.168
RW, 100% user accuracy	0.164
Y, 100% user accuracy	0.195
Best Interactive Search	0.303
Top-10-Search-Average	0.228
CMU Automated Search	0.079

Drawing fewer top-ranked shots from across more concepts in general does not work well. Only 2 topics show benefit from drawing from more than 2 concepts: topics 173 "emergency vehicles" and 176 "escort prisoner." In part this is due to LSCOM-Lite being just that, a "light" set without the ontological framework making it difficult to find realistic topics that cut across more than a few topics equally. In part this is due to some topics being covered so well by one or at most a pair of concepts that introducing additional concepts only reduces the relevance in the candidate set.

From Table 5, it appears that presenting the very best shots from 1, 2, or 4 concepts all produce quite good MAP near to or better than the top 10 TRECVID 2006 search systems. Even the searches without an oracle can approach reasonable interactive performance, much better than a fully automated search. The numbers are overly optimistic in retrospect, but the full post-hoc analysis is kept in this paper as a caution to others conducting analyses with simulated rather than actual users. In our case, we have empirical data from which to gauge whether the 100% user precision and recall are realistic, the subject of the next section.

4. Replacing perfect user performance with "realistic" empirically based data

As with our limit of 2000 on the size of the shot set to be considered, we again turn to empirical data to help push the reported concept analysis further. We have metrics on the number of shots collected by 3 users across all 24 TRECVID 2006 topics each [6], in line with the empirical data collected through additional user studies conducted on earlier TRECVID data sets [2, 3, 7, 17]. For each topic, the number of shots reviewed, number of shots marked as definitely addressing the topic (primary shot set), number of shots marked as possibly addressing the topic (secondary shot set), and number of shots overlooked as not addressing the topic (overlooked shot set) are tracked. In our Informedia interface, we have the capability to better rank tagged results by keeping the "maybe, not quite sure" category of secondary shots. Table 6 summarizes precision for the collected data for the three users. When these users overlook a shot, 97.2% of the time that shot is indeed not relevant to the topic at hand. When these users mark a shot as definitely relevant, it does have significantly higher precision over the "perhaps relevant" secondary set. Alas, the precision of the primary set is still far from perfect, on average 74% given the difficulty in resolving correctness of some shots to topics based often only on quick inspection of a thumbnail image. This data indicates that our assumption on 100% user precision for Figure 1 and Table 5 is unlikely given the nature of the TRECVID 2006 topics.

Table 6. Precision for 3 users averaged across the 24 TRECVID 2006 topics for 3 collected shot sets: primary (user marked as correct), secondary (user marked as maybe) and overlooked (user skipped).

User	Primary	Secondary	Overlooked
A	0.799	0.336	0.974
B	0.790	0.279	0.977
C	0.635	0.241	0.966

We considered the mythical user for Figure 1's data to have perfect recall. To derive a more realistic and practical number for recall from the set of 2000, we look to the number of correct shots that were in the three users' primary plus secondary plus overlooked shot sets. If the user were given X shots to inspect with X on average near 2000, and R of them are relevant, how many of the R will the user spot? How many of the correct shots in primary plus secondary plus overlooked were actually tagged by the user for inclusion in the ***primary*** set? We could have relaxed the recall constraint and given the user credit for putting relevant shots in *either* of the ***primary*** **or** ***secondary*** sets, but we opted to be stricter and count only the primary set, producing a lower recall number across the users and topics. On average, user A placed 53.7% of the correct shots (post-hoc judgment via

NIST pooled truth) for a topic into A's primary set. User B placed 44.0% and user C 36.4%. For example, for topic 175 "enter/exit vehicles" user A had 32 relevant shots available in the full set of primary plus secondary plus overlooked shots reviewed by A. The user marked 27 as primary (definitely relevant) of which 24 actually were judged as relevant in the NIST pooled truth (see [1] for discussion on "pooled truth" for TRECVID 2006). User A's precision on this topic was 24/27 (88.9%) and his recall from the set of shots he inspected was 24/32 = 75%. The numbers vary by topic, with some topics scoring high precision and recall and others being more difficult. Rather than lose those differences by applying an average throughout, we keep the user-derived precision and recall numbers on each of the 24 topics, shown in Figure 2.

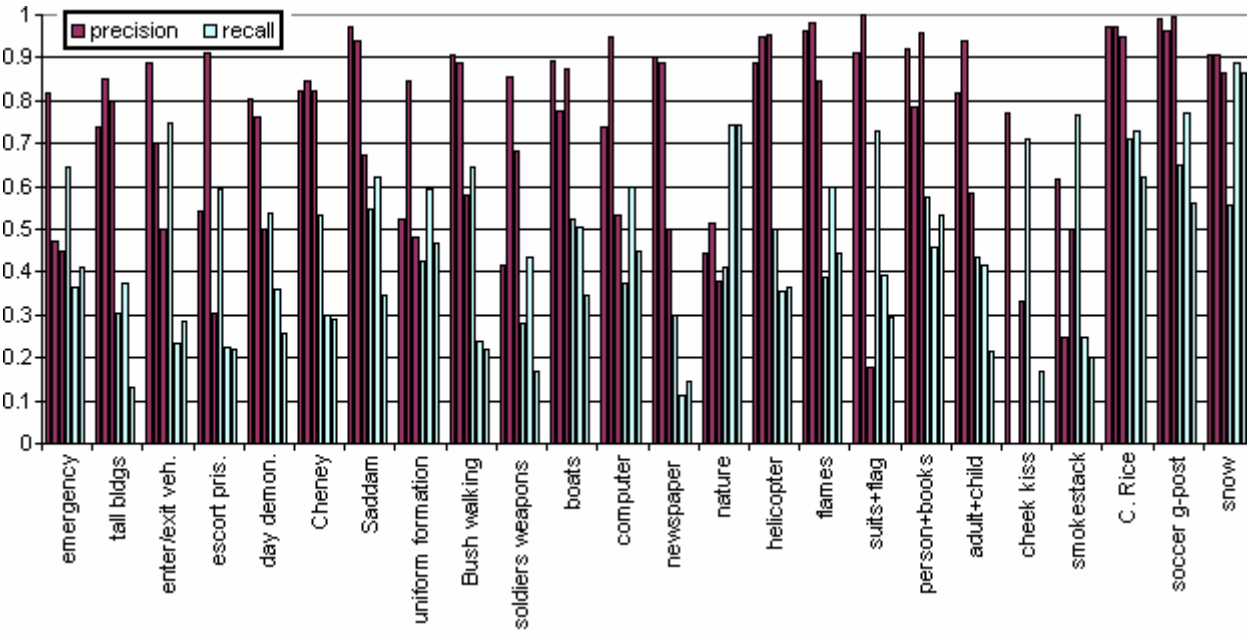

Figure 2. Precision and recall numbers from users A, B, C used to derive empirical numbers for topics, e.g., for topic "adult plus child", empirical precision and recall values are 0.78 and 0.36.

Armed with this empirical data, we generated a follow-up to Table 5, shown in Table 7.

Table 7. MAP across various TRECVID 2006 runs with perfect and average (from A, B, C) user.

Run	MAP
1at2000, 100% user	*0.265*
1at2000, average user	0.112
2at1000, 100% user	*0.248*
2at1000, average user	0.108
4at500, 100% user	*0.220*
4at500, average user	0.095
BF1, average user	0.074
RW, average user	0.083
Y, average user	0.095
Best Interactive Search	0.303
Top-10-Search-Average	0.228
CMU Automated Search	0.079

Conclusions drawn earlier about the relative effectiveness of 1 concept vs. 2 or more still stand, as the reduced but more realistic performance numbers in Table 7 do not change the relative ordering in the "1at2000" – 2 – 4 rows from what is shown in Table 5. However, they do significantly reduce the MAP. Now, query-by-concept by itself as an interaction strategy, even if ideally informed by a concept selection strategy that always picks the best concept for a topic, still falls well below the performance of the best interactive system using query-by-text, query-by-image, and query-by-concept.

5. Conclusions and future work

When watered down by true user-based performance metrics, the optimal MAP numbers of Section 3 no longer hold true, and differences between considered concept selection strategies begin to soften because of introduced user inaccuracies. For TREC-VID 2006, many of the topics were made more difficult so that simple thumbnail review from storyboards would not produce high recall. The

investigations outlined here, from the use of oracle-based concept recommendation to user-driven selection strategies needing no oracle but good topic recognition, should be continued with other topic sets, ideally derived from real user communities reflecting true video corpus needs. The investigations should also continue with a dramatically larger LSCOM concept set, rather than the small LSCOM-Lite set. With hundreds of concepts, it will be interesting to see whether realistic topics are still best addressed with one or two concepts, as found here, or if the best shots from multiple concepts will result in better topic performance. An alternative to the empirical strategies for selecting concepts and shots (i.e., BF1, RW, and Y) is to use probabilistic local content analysis (pLCA) as described in [18], however, we would not expect the results to change much relative to the trends reported here.

The user interface remains important, as the difference between Section 3 and Section 4 shows. Users are not perfect, so post-hoc analyses are cautioned against assuming 100% user performance and should bring in empirical measures as well. If the interface can enable not only the efficient review of 1000s of shots, but the effective review so that user precision and recall can inch closer to the idealistic 100% recall and precision from a scanned set of 2000 shots, then users will be able to achieve great search utility. In fact, with such high levels of precision and recall against a system-provided set from just query-by-concept, the user will be able to achieve performance levels previously reached only by the top-scoring systems using query-by-text, by-image, and by-concept.

6. Acknowledgments

This work is supported by the National Science Foundation under Grant No. IIS-0205219. Our thanks to NIST and the TRECVID organizers for enabling this video retrieval evaluation work. Details about Informedia research and the full project team can be found at www.informedia.cs.cmu.edu.

7. References

[1] Kraaij, W., Over, P., Ianeva, T., and Smeaton, A., "TRECVID 2006 - An Introduction", *TRECVID Online Proceedings,* http://www-nlpir.nist.gov/projects/tvpubs/ in location tv6.papers/tv6intro.pdf, Nov. 2006.

[2] Hauptmann, A., and Christel, M., "Successful Approaches in the TREC Video Retrieval Evaluations", *Proc. ACM Multimedia* (New York, Oct. 2004), pp. 668-675.

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

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

[5] Snoek, C., Worring, M., et al., "A Learned Lexicon-Driven Paradigm for Interactive Video Retrieval", *IEEE Trans. Multimedia* 9(2), Feb. 2007, pp. 280-292.

[6] Christel, M., and Yan, R., "Merging Storyboard Strategies and Automatic Retrieval for Improving Interactive Video Search", *Proc. Image and Video Retrieval (CIVR)* (Amsterdam, The Netherlands, July 2007).

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

[8] Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., and Jain, R., "Content based image retrieval at the end of the early years", *IEEE Trans. Pattern Analysis and Machine Intelligence* 22(12), 2000, pp. 1349-1380.

[9] Markkula, M. and Sormunen, E., "End-user searching challenges indexing practices in the digital newspaper photo archive", *Information Retrieval* 1(4), 2000, pp. 259-285.

[10] Moënne-Loccoz, N., et al., "Managing Video Collections at Large", *ACM Proc. Workshop on Computer Vision meets Databases* (Paris, June 2004), pp. 59-66.

[11] Naphade, M., et al., "Large-Scale Concept Ontology for Multimedia", *IEEE MultiMedia* 13(3), 2006, pp. 86-91.

[12] Soo, V.-W., Lee, C.-Y., Li, C.-C., Chen, S.L., and Chen, C., "Automated Semantic Annotation and Retrieval Based on Sharable Ontology and Case-based Learning Techniques", *Proc. ACM/IEEE JCDL* (Houston, May 2003), pp. 61-72.

[13] Yang, M., et al., "The relative effectiveness of concept-based versus content-based video retrieval", *Proc. ACM Multimedia* (New York, Oct. 2004), pp. 368-371.

[14] Cao, Y. et al., "Intelligent Multimedia Group of Tsinghua University at TRECVID 2006", Nov. 2006, *TRECVID Online Proceedings*.

[15] Campbell, M., et al., "IBM Research TRECVID-2006 Video Retrieval System", *TRECVID Proc.,* http://www-nlpir.nist.gov/projects/tvpubs/tv6.papers/ibm.pdf, Nov. 2006.

[16] Christel, M., Naphade, M., Natsev, A., and Tesic, J., "Assessing the Filtering and Browsing Utility of Automatic Semantic Concepts for Multimedia Retrieval," *CVPRW '06: Proc. Conf. Computer Vision and Pattern Recognition Workshop* (New York, June 2006).

[17] Christel, M., "Establishing the Utility of Non-Text Search for News Video Retrieval with Real World Users," *Proc. ACM Multimedia* (Augsburg, Germany, Sept. 2007).

[18] Yan, R., "Probabilistic Models for Combining Diverse Knowledge Sources in Multimedia Retrieval", PhD Thesis, Language Technologies Institute, School of CS, Carnegie Mellon University, Pittsburgh, PA, USA, 2007.