User Strategies in Video Retrieval: a Case Study

L. Hollink[1], G.P. Nguyen[2], D.C. Koelma[2], A.Th. Schreiber[1], M. Worring[2]

[1] Business Informatics, Free University Amsterdam. {hollink,schreiber}@cs.vu.nl
[2] ISIS, University of Amsterdam. {giangnp,koelma,worring}@science.uva.nl

Abstract. In this paper we present the results of a user study that was conducted in combination with a submission to TRECVID 2003. Search behavior of students querying an interactive video-retrieval system was analyzed. 242 Searches by 39 students on 24 topics were assessed. Questionnaire data, logged user actions on the system, and a quality measure of each search provided by TRECVID were studied. Analysis of the results at various stages in the retrieval process suggests that retrieval based on transcriptions of the speech in video data adds more to the average precision of the result than content-based retrieval. The latter is particularly useful in providing the user with an overview of the dataset and thus an indication of the success of a search.

1 Introduction

In this paper we present the results of a study in which search behavior of students querying an interactive video-retrieval system was analyzed. Recently, many techniques have been developed to automatically index and retrieve multimedia. The Video Retrieval Track at TREC (TRECVID) provides test collections and software to evaluate these techniques. Video data and statements of information need (topics) are provided in order to evaluate video-retrieval systems performing various tasks. In this way, the quality of the systems is measured. However, these measures give no indication of user performance. User variables like prior search experience, search strategies, and knowledge about the topic can be expected to influence the search results. Due to the recent nature of automatic retrieval systems, not many data are available about user experiences. We argue that knowledge about user behavior is one way to improve performance of retrieval systems. Interactive search in particular can benefit from this knowledge, since the user plays such a central role in the process.

We study information seeking behavior of users querying an interactive video-retrieval system. The study was conducted in combination with a submission to TRECVID 2003 [1]. Data were recorded about user characteristics, user estimations of the quality of their search results, familiarity of users with the topics, and actions performed while searching. The aim of the study was to investigate the influence of the recorded user variables on the average precision of the search results. In addition, a categorization was made of the 24 topics that were provided by TRECVID. The categories show differences in user behavior and average precision of the search results.

2 Research Questions

To gain knowledge about how user-related factors affect search in a state-of-the-art video-retrieval system, we record actions that users take when using such a system. In particular, we are interested in which actions lead to the best results. To achieve an optimal search result, it is important that a user knows when to stop searching. In this study we therefore measure how well users estimate the precision and recall of their search.

It is possible that different topics or categories of topics lead to different user strategies and differences in the quality of the results. We compare the search behavior and search results of categories of topics. In sum, the main questions in the study are:

1. What search actions are performed by users and which actions lead to the best search results?
2. Are users able to estimate the success of their search?
3. What is the influence of topic type on user actions and search results?

3 The ISIS Video Retrieval System

The video-retrieval system on which the study was performed was built by the Intelligent Sensory Information Systems (ISIS) group at the University of Amsterdam for the interactive video task at TRECVID. For a detailed description of the system we refer to [1].

The search process consists of four steps: indexing, filtering, browsing and ranking. Indexing is performed once off-line. The other three steps are performed iteratively during the search task. The aim of the *indexing* step is to provide users with a set of high-level entry points into the dataset. We use a set of 17 specific concept detectors developed by CMU for the TRECVID, such as female speech, aircraft and newsSubjectMonologue. We augment the high-level concepts by deriving textual concepts from the speech recognition result using Latent Semantic Indexing (LSI). Thus we decompose the information space into a small set of broad concepts, where the selection of one word from the concept reveals the complete set of associated words also.

For all keyframes in the dataset low-level indexing is performed by computing the global Lab color histograms. To structure these low-level visual descriptions of the dataset, the whole dataset is clustered using k-means clustering with random initialization. The k in the algorithm is set to 143 as this is the number of images the display will show to the user. In summary, the off-line indexing stage results in three types of metadata associated with each keyframe: (1) the presence or absence of 17 high-level concepts, (2) words occurring in the shot extended with associated words and (3) a color histogram.

After indexing, the interactive process starts. Users first *filter* the total corpus of video by using the indexing data. Two options are available for filtering: selecting a high-level concept, and entering a textual query that is used as a concept. These can be combined in an 'and' search, or added in an 'or' search.

The filtering stage leads to an active set of shots represented as keyframes, which are used in the next step, *browsing*. At this point in the process it is assumed that the user is going to select relevant keyframes from within the active set. To get an overview of the data the user can decide to look at the clustered data, rather than the whole dataset. In this visualization mode, the central keyframe of each cluster is presented on the screen, in such a way that the distances between keyframes are preserved as good as possible. The user interface does not play the shots as clips since too much time would be spend on viewing the video clips.

When the user has selected a set of suitable images, the user can perform a *ranking* through query-by-example using the color histograms with Euclidean distance. The closest matches within the filtered set of 2,000 shots are computed, where the system alternates between the different examples selected. The result is a ranked list of 1,000 keyframes.

4 Methods

We observed search behavior of students using the video-retrieval system described in Sect. 3. The study was done as an addition to a submission to TRECVID. Apart from the search results that were collected and submitted to TRECVID, additional user-related variables were collected.

For the TRECVID 24 topics had to be found in a dataset consisting of 60 hours of video from ABC, CNN and C-SPAN. 21 Groups of students (18 pairs and 3 individuals) were asked to search for 12 topics. The topics were divided into two sets of 12 (topics 1-12 and topics 13-24) and assigned a set to each student pair. For submissions to TRECVID the time to complete one topic was limited to 15 minutes. Prior to the study the students received a three-hour training on the system. Five types of data were recorded:

Entry Questionnaire Prior to the study all participants filled in a questionnaire in which data was acquired about the subject pool: gender, age, subject of study, year of enrollment, experience with searching.

Average Precision Average precision (AP) was used as the measure of quality of the results of a search. AP is the average of the precision value obtained after each relevant camera shot is encountered in the ranked list [1]. Note that AP is a quality measure for *one search* and not the mean quality of a group of searches. AP of each search was computed with a ground truth provided by TRECVID. Since average precision fluctuates during the search, we recorded not only the average precision at the end of the search but also the maximum average precision during the search.

Logfiles Records of user actions on the system were made containing the following data about each search: textual queries, high-level features used, type of query ('and' or 'or'), number of images selected, duration of the search. These data were collected at two points in time: at the end of the search and at the point at which maximum average precision was reached. The logfile data are used to answer the first research question.

Topic Questionnaire After each search the participants answered 5 questions about the search: 1. Are you familiar with this topic? 2. Was it easy to get started on this search? 3. Was it easy to do the search on this topic? 4. Are you satisfied with your search results? 5. Do you expect that the results of this search contain a lot of non-relevant items (low precision)? All questions were answered on a 5-point scale (1=not at all, 5=extremely). The resulting data were used as input for answering the second research question.

Exit Questionnaire After the study all participants filled in a short questionnaire containing questions about the user's opinion of the system and the similarity between this type of search and the searches that they were used to perform.

To answer the third research question, the topics were categorized using a framework that was designed for a previous study [2]. The framework combines different methods to categorize image descriptions (e.g [3] and [4]) and divides queries into various levels and classes. For the present study we used only those distinctions that we considered relevant to the list of topics provided by TRECVID (Table 1): "general" vs. "specific" and "static" vs. "dynamic". Other distinctions, such as "object" vs. "scene", were not appropriate for the topic list since most topics contained descriptions of both topics and scenes.

Table 1. Summary of topics, categorized into general and specific and into dynamic and static. See http://www.cs.vu.nl/~laurah/trec/topics.html for topic details.

Class	General	Specific
Static	18: a crowd in urban environment	09: the mercedes logo
	16: road with vehicles	25: the white house
	14: snow-covered mountains	07: tomb of the unknown soldier
	13: flames	17: the sphinx
	01: aerial view of buildings	24: Pope John Paul II
	10: tank	04: Yassar Arafat
	22: cup of coffee	20: Morgan Freeman
	23: cats	15: Osama bin Laden
	06: helicopter	19: Mark Souder
Dynamic	05: airplane taking off	02: basketball passing down a hoop
	12: locomotive approaching you	03: view from behind catcher while pitcher is throwing the ball
	08: rocket taking off	
	11: person diving into water	

5 Subjects

The subjects participating in the study were 39 students in Information Science who enrolled in the course Multimedia Retrieval at the University of Amsterdam. The number of years of enrollment at the university was between 1 and 8 (mean = 3.5). Two were female, 37 male. Ages were between 20 and 40 (mean=23.4).

Before the start of this study, we tested the prior search experience of the subjects in a questionnaire. All subjects answered questions about frequency of use and experience with information retrieval systems in general and, more specifically, with multimedia retrieval systems. It appeared that all students searched for information at least once a week and 92 % had been searching for two years or more. All students searched for multimedia at least once a year, and 65 % did this once a week or more. 88 % of the students had been searching for multimedia for at least two years. This was tested to make sure that prior search experience would not interfere with the effect of search strategies on the results. We did not find any evidence of a correlation between prior search experience and strategy, nor between prior search experience and search results. The lack of influence of search experience can in part be explained from the fact that the system was different from search systems that the students were used to. All but three students indicated in the exit questionnaire that the system was not at all similar to what they were used to. All students disagreed with or were neutral to the statement that the topics were similar to topics they typically search for. Another possible reason for the absence of an effect of prior search experience is the three-hour training that all students had received before the study.

The subjects indicated a high familiarity with the topics. Spearman's correlation test indicated a relationship between familiarity and average precision only within topics 10 and 13. We do not consider this enough evidence that there is in fact a relationship.

6 Results

The data were analyzed on the level of individual searches. A search is the process of one student pair going through the three interactive stages of the system for one topic. 21 Groups of students searched for 12 topics each, resulting in 252 searches. After exclusion of searches that were not finished, contained too much missing data, or exceeded the by TRECVID imposed maximum of 15 minutes, 242 searches remained.

User actions. In Table 2 descriptives are presented of the variables recorded in the logfiles. It shows that a search took approximately 8 minutes; 9 images were selected per search; high-level features were hardly used; or-search was used more than and-search.

The mean average precision at the end of a search was 0.16. *Number of selected images* was the most important variable to explain the result of a search. This can be explained by the fact that each correctly selected image adds at least one relevant image to the result set. The contribution of the ranking to the result was almost negligibly small; change in AP caused by the ranking step had a mean of 0.001 and a standard deviation of 0.032. *Number of selected images* was not correlated to *time to finish topic*, *number of features*, or *type of search*.

There was no correlation between *time to finish topic* and average precision, nor between *type of search* and average precision. *Number of high-level features*

Table 2. User actions in the system at the moment of maximum AP and at the end of the search

		Max				End			
	N	Min.	Max.	Mean	St.D.	Min.	Max.	Mean	St.D.
Time (sec.)	242	0	852	345	195	6	899	477	203
No. of images selected	242	0	30	8.47	7.01	0	30	9.07	7.06
No. of high-level features	240	0	5	0.50	0.84	0	17	0.59	1.39
'And' or 'Or' search	240	And:75 Or:165				And:82 Or:158			

had a negative influence on the result. This is depicted in Fig. 1. The number of uses per features was too low to draw conclusions about the value of each feature. We can conclude, however, that selection of more than one feature leads to low average precision. To give an indication of the quality of the features that were used by the students, Table 3 shows the frequency of use and the mean average precision of the features. Only searches in which a single feature was used are included.

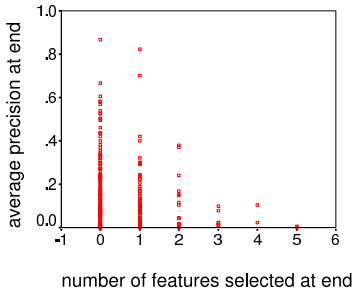

Fig. 1. Scatterplot of number of selected features and AP at the end of the search. One case with 17 features and AP of 0.027 is left out of the plot.

User prediction of search quality. In the topic questionnaire we collected opinions and expectations of users on a particular search. All questions measure an aspect of the user's estimation of the search. For each question it holds that a high score represents a positive estimation, while a low score represents a negative estimation. Mutual dependencies between the questions complicate conclusions on the correlation between each question and the measured average precision of a search. Therefore, we combined the scores on the 4 questions into one variable, using principal component analysis. The new variable that is thus created represents the combined user estimation of a search. This variable explains 70 % of the variance between the cases. Table 4 shows the loading of

Table 3. High-level features: mean average precision and standard deviation.

Feature	N	Mean AP	St.d.	Feature	N	Mean AP	St.d.
Aircraft	5	0.09	0.05	People	3	0.13	0.15
Animal	5	0.17	0.06	PersonX	7	0.14	0.16
Building	2	0.30	0.00	PhysicalViolence	0	.	.
CarTruckBus	4	0.11	0.03	Road	3	0.06	0.04
FemaleSpeech	0	.	.	SportingEvent	9	0.08	0.03
NewsSubjectFace	1	0.24	.	Vegetation	1	0.13	.
NewsSubjectMonologue	1	0.70	.	WeatherNews	0	.	.
NonStudioSetting	4	0.15	0.13	ZoomIn	1	0.08	.
Outdoors	15	0.17	0.20				

each question on the first principal component. Pearson's correlation test showed a relationship between combined user estimation and actually measured average precision. (Pearson's correlation coefficient (Pcc) = 0.298, $\alpha = 0.01$). This suggests that users are indeed able to estimate the success of their search.

Table 4. Principal Component Analysis

Questionnaire item	Component 1
easy to start search	0.869
easy to do search	0.909
satisfied with search	0.874
expect high precision	0.678

Another measure of user estimation of a search is the difference between the point where maximum precision was reached and the point where the user stopped searching. As mentioned in Sect. 6, the mean time to finish a search was 477 seconds, while the mean time to reach maximum average precision was 345 seconds. The mean difference between the two points in time was 128 seconds, with a minimum of 0, a maximum of 704 and a standard deviation of 142 seconds. This means that students typically continued their search for more than two minutes after the optimal result was achieved. This suggests that even though students were able to estimate the overall success of a search, they did not know when the best results were achieved within a search. A correlation between combined user estimation and time-after-maximum-result shows that the extra time was largest in searches that got a low estimation (Pcc = -0.426, $\alpha = 0.01$). The extra 2 minutes did not do much damage to the precision. The mean average precision of the end result of a search was 0.16, while the mean maximum average precision of a search was 0.18. The mean difference between the two was 0.017, with a minimum of 0, a maximum of 0.48 and a standard deviation of 0.043.

Topic type. Table 5 shows that "specific" topics were better retrieved than "general" topics. The results of "static" topics were better than the results of "dynamic" topics. These differences were tested with an analysis of variance. The differences are significant far beyond the 0.01 α-level. We did not find any evidence that user actions were different in different categories.

Table 5. Mean AP of topics types, and ANOVA results

Mean AP	Static	Dynamic	Total
General	0.12	0.10	0.11
Specific	0.27	0.08	0.22
Total	0.19	0.10	0.16

ANOVA results	SS	df	MS	F	Sig.
Between Groups	0.426	1	0.426	18.109	0.000
Within groups	5.648	240	0.024		
Total	6.074	241			

The change in AP caused by the ranking step was positive for general topics (mean = 0.005), while negative for specific topics (mean = - 0.004). For general topics we found a correlation between *change in AP* and *AP at the end of the search* (Pcc = 0.265, $\alpha = 0.004$), which was absent for specific topics.

7 Discussion

Different types of topics result in differences in the quality of the search results. Results of "specific" topics were better than results of "general" topics. This suggests that indexing and filtering are the most important steps in the process. These steps are based on text retrieval, where it is relatively easy to find uniquely named objects, events or people. In content-based image retrieval on the other hand, and especially when the image is concerned as a whole, it is difficult to distinguish unique objects or people from other items of the same category. We are planning to upgrade the system so that regions within an image can be dealt with separately. Results of "static" topics were better than results of "dynamic" topics. This can be explained by the fact that the system treats the video data in terms of keyframes, *i.e.,* still images.

From the recorded user actions, *number of selected images* is by far the most important for the result. This is mainly caused by the addition of correctly selected images to the result set. The contribution of the ranking step to the average precision was almost negligibly small. We conclude from this that the main contribution of content-based image retrieval to the retrieval process is visualization of the dataset which gives the user the opportunity to manually select relevant keyframes. The visualization of the data set also gives the user an overview of the data and thus an indication of the success of the search. The results of the study show that users can estimate the success of a search quite well, but do not know when the optimal result is reached within a search.

This study reflects user behavior on one particular system. However, the results can to a certain extent be generalized to other interactive video-retrieval systems. The finding that "specific" topics are better retrieved than "general"

topics is reflected by the average TRECVID results. The fact that users do not know when to stop searching is a general problem of category search [5], where a user is searching for shots belonging to a certain category rather than for one specific shot. One solution to this problem is providing the user with an overview of the dataset. Future research is needed to compare the effectiveness of different types of visualization.

One of the reasons for this study was to learn which user variables are of importance for video retrieval, so that these variables can be measured in a future experiment. The most discriminating variable in the study proved to be the *number of selected images*. Further research is needed in which the optimal number of examples in a query-by-example is determined, taking in account the time spent by a user. In addition, future research is needed in which the four steps in the system are compared. In an experimental setting text-based retrieval and content-based retrieval can be compared. It would also be interesting to compare the results of an interactive video retrieval system to sequential scanning of shots in the data set for a fixed amount of time.

One of the results was that prior experience with searching and familiarity with the topic do not affect the quality of the search results. The latter seems to indicate that background knowledge of the searcher about the topic is not used in the search process. Some attempts to include background knowledge into the process of multimedia retrieval are made (see for example [6, 7]). We would be interested to see how these techniques can be incorporated in an interactive video retrieval system.

References

1. M.Worring, G.P.Nguyen, L.Hollink, J.Gemert, D.C.Koelma: Interactive search using indexing, filtering, browsing, and ranking. In: Proceedings of TRECVID. (2003)
2. L.Hollink, A.Th.Schreiber, Wielinga, B., M.Worring: Classification of user image descriptions. International Journal of Human Computer Studies (2004) To Appear.
3. Armitage, L., Enser, P.: Analysis of user need in image archives. Journal of Information Science **23** (1997) 287–299
4. Jorgensen, C.: Attributes of images in describing tasks. Information Processing and Management **34** (1998) 161–174
5. Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000)
6. L.Hollink, A.Th.Schreiber, J.Wielemaker, B.Wielinga: Semantic annotation of image collections. In: Proceedings of the K-CAP 2003 Workshop on Knowledge Markup and Semantic Annotation, Florida, USA (2003)
7. A.Jaimes, B.L.Tseng, J.R.Smith: Modal keywords, ontologies, and reasoning for video understanding. In E.M.Bakker, ed.: CIVR. Volume 2728 of LNCS. (2003)