Assessing user behaviour in news video retrieval

L. Hollink, G.P. Nguyen, D.C. Koelma, A.Th. Schreiber and M. Worring

Abstract: The results of a study are presented, in which people queried a news archive using an interactive video retrieval system. 242 search sessions by 39 participants on 24 topics were assessed. Before, during and after the study, participants filled in questionnaires about their expectations of a search. The questionnaire data, logged user actions on the system, queries formulated by users, and a quality measure of each search were studied. The results of the study show that topics concerning 'specific' people or objects were better retrieved than topics concerning 'general' objects and scenes. Users were able to estimate the overall quality of a search but did not know when the optimal result was reached within the search process. 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 image retrieval based on low-level visual features. 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. Based on the results, implications for the design of user interfaces of video retrieval systems are discussed.

1 Introduction

In this paper we study information seeking behaviour of users searching in a collection of broadcast news video. Large collections of broadcast material are maintained at broadcasting stations and at archiving organisations such as 'The Netherlands Institute for Sound and Vision' and the 'Institut National de l'Audiovisuel'. In recent years, these archives have been queried by a broad user group, including broadcasters, documentary makers, researchers and students. However, access to broadcast video is still difficult and too often a time consuming process [1].

Many techniques have been developed to automatically index and retrieve multimedia. The TREC Video Retrieval Evaluation (TRECVID) [Note 1] 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 performance of the systems is measured. However, these measures give no indication of how user behaviour and user characteristics affect the performance of retrieval systems. User variables like prior search experience, search actions, and knowledge about the topic can be expected to influence the search results. Owing to the recent nature of automatic retrieval systems, not many data are available about user experiences. We argue that knowledge about user behaviour 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. Studies have been done to measure usability of interactive retrieval systems (e.g. [2]) and effectiveness of different components of these systems ([3]). In this paper we investigate the still unclear impact of user behaviour and user characteristics on the performance of interactive retrieval systems.

We participated in the interactive search task of TRECVID and explored user behaviour on a state-of-the-art interactive news video retrieval system [4]. The TRECVID collection consists of 60 hours of video from ABC, CNN and C-SPAN, and 24 topics. News data can, in theory, contain every theme in the world, which complicates the retrieval process. However, this broadness also makes it a valuable test collection, since the results will be applicable to a wide range of collections. Within this broad context we focus on category search: a user is searching for shots belonging to a certain category rather than for one target shot.

In this study we record data about user characteristics, familiarity of users with topics, queries formulated by users, and actions that users take when using the system. In particular, we are interested in which actions lead to the best results. To achieve an optimal search result, a user needs to have a good overview of the contents of the collection. This will give the user an idea of the recall and precision of a search, and will aid the user during the search process in deciding whether a continuation of the search is likely to yield new and better results. Therefore, in this study we measure how well users estimate the quality of their search.

In addition, a categorisation was made of the 24 topics. It is possible that different categories of topics lead to different user actions and differences in the quality of the

IEE Proceedings online no. 20045187

doi: 10.1049/ip-vis:20045187

Paper first received 24th September 2004 and in revised form 13th June 2005

L. Hollink and A.Th. Schreiber are with the Section Business Informatics, Free University Amsterdam, De Boelelaan 1081a, 1081HV Amsterdam, The Netherlands

G.P. Nguyen, D.C. Koelma and M. Worring are with the Intelligent Sensory Information Systems Group, University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands

E-mail: hollink@cs.vu.nl

Note 1: http://www-nlpir.nist.gov/projects/trecvid/

results. We compare search behaviour and search results of categories of topics.

In summary, 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 category on user actions and search results?

This study is an extension of previous work [5].

2 The interactive video retrieval system

2.1 Indexing of the video data

Prior to user interaction, the whole collection of video data is indexed in order to provide the user with high-level entry points into the dataset. First, we derive high-level textual concepts from the automatic speech recognition (ASR) result [6] using latent semantic indexing (LSI) [7]. To that end, we construct a vector space by taking all words found in the ASR results of all videos in the collection. We then perform stopword removal using the SMART's English stoplist. This results in a 18,117-dimensional vector space. Using LSI the vector space is reduced to 400 dimensions. 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.

Second, we use 17 high-level concept detectors developed by Carnegie Mellon University (CMU) for the TRECVID [8], ranging from generic ones like outdoors to more specific ones like physical violence. The quality of the detectors ranges from poor to good.

In addition, for all keyframes in the dataset we perform low-level indexing by computing the global Lab colour histograms using 32 bins for each channel. To structure these low-level visual descriptions, the whole dataset is clustered using k-means clustering with random initialisation. The k in the algorithm is set to 143 as this is the number of images our display will show to the user.

2.2 User interaction with the system

User interaction with the system consists of two steps: (1) filtering of the complete data set into a smaller 'active set', and (2) browsing through the active set. A user enters the system with an information need. In TRECVID, statements of information need are statements like 'find shots of an airplane taking off', or 'find shots of the Sphinx'. A typical session on the system starts with a user entering a textual query (Fig. 1). The user then chooses between 'exact search' (without LSI) or 'concept search' (with LSI). By default the system is set to 'concept search' (Fig. 1). In addition, the user can indicate the desired presence or absence of each of the 17 high-level concepts. Users can combine the two query mechanisms using an AND function (but this usually leads to very small sets and low recall) or an OR function, where the ranked result is an alternation between the results obtained for the selected query specification mechanisms. The default value is OR. The two mechanisms together produce a ranked list of shots, the active set, that is used in the subsequent browsing step. We restrict the active set to contain 2000 shots maximum, leading to approximately 4000 keyframes.

In the browsing step keyframes from the active set are displayed to the user. Browsing requires a visualisation mechanism that on the one hand provides an overview of the dataset, while showing sufficient detail on the other. Furthermore, the visualisation should give the user an insight in the structure of the dataset. The system supports the user with an array-based (Fig. 2) and a similarity-based (Fig. 3) visualisation. When the user points to a thumbnail of a keyframe, a full size image and text associated with the shot are shown on the right side of the screen. Sequential keyframes in the video from which a keyframe is selected, are presented at the bottom of the screen (Fig. 3).

The user can now select relevant example keyframes from within the active set. When the user has selected a set of examples, he or she can click the 'feedback' button in order to obtain a ranked result list of images from the active set. Ranking of the active set is based on query-by-example (QBE) where similarity of two keyframes is defined by the Euclidean distance of the two Lab histograms. In the result the closest matches with the example images are computed, where the system alternates between the different examples selected. The user-selected example images are placed in the highest ranks of the result list. To allow for easy comparison between systems, we follow the TRECVID custom by always letting the result of a search consist of a ranked list of 1000 items.

If the resulting ranked list of keyframes is not satisfying, the user can decide to go back to the filtering stage and change the query, or to continue browsing for relevant examples and perform a new ranking. The re-ranked set is again visualised

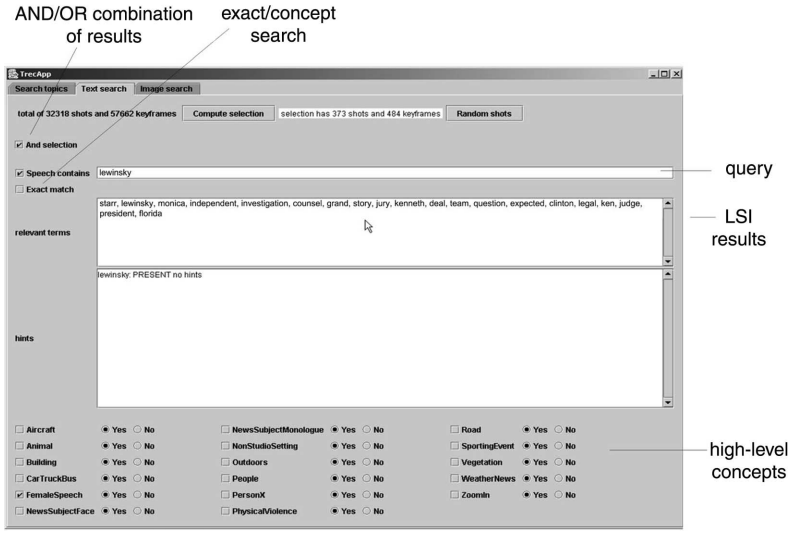

Fig. 1 *Screen shot of the GUI used for query entering*

Fig. 2 *Screen shot of the GUI used for browsing with array-based visualisation*

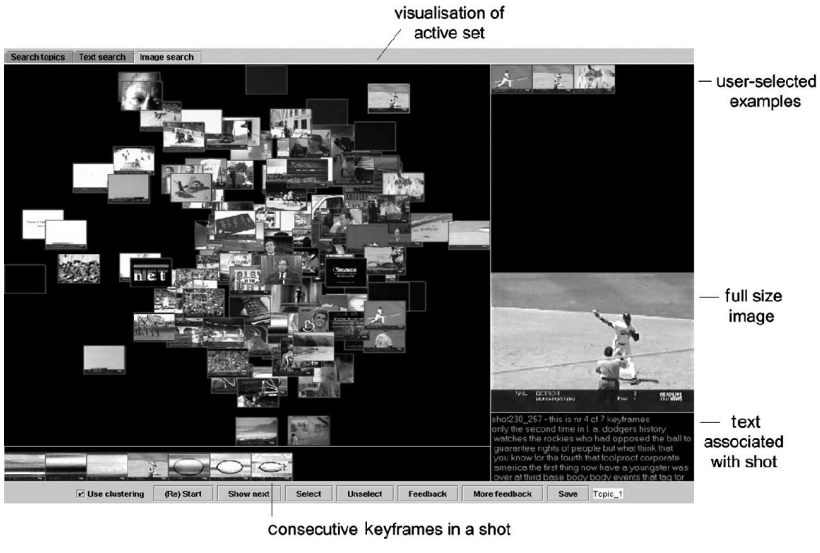

Fig. 3 *Screen shot of the GUI used for browsing with similarity-based visualisation*

as in Figs. 2 or 3, enabling the discovery of new relevant keyframes. The process of querying, browsing and (re)ranking continues until the user is satisfied and saves the results.

2.3 Comparison to other systems

User actions on a system always take place within the bounds of the user interface of the system. Comparing the user interface of the current system to similar systems gives an indication of how much of the user behaviour is specific to the current system, and how much is likely to be typical to interactive video retrieval systems in general. In the TRECVID 2003 conference, 12 systems participated in the interactive search task [9]. We compare the features of our system to five well performing systems from IBM ([10], Carnegie Mellon University [8], Dublin City University [11], Imperial College London [12], and the University of Oulu [13]). Some of the systems come in multiple forms (e.g. a text only system and a combined text/image system). In these cases we look at the most extensive variant. We do not seek to do justice to all design decisions that have been made in these systems. Instead, we try to give a short overview of the major functionalities of the user interfaces. We will not go into the underlying techniques that are hidden from the user, even though these techniques are no doubt of major importance for differences in performance.

All systems provide the user with a field in which free text queries can be typed, and all systems use ASR results to process textual queries. Also, all systems have a mechanism to let users select positive example images while browsing. Positive examples can be added to the query and will appear high in the result list. The system from Dublin adds not only the keyframe, but also the associated text to the query when a keyframe is selected as a positive example. The IBM system lets users select negative examples ([10]). Two systems ([8, 13]) use the high-level concepts from the TRECVID feature task (e.g. car/truck/bus, female speech, outdoor) to filter the dataset in a similar fashion to our system. None of the systems uses LSI to extend textual queries.

There are different ways to combine the components of a query (text, example images, high-level concepts). Two systems let users adjust the weights of textual and image queries ([10, 11]). The system from Oulu lets users switch text-based, image-based and high-level-concept-based searches on or off. In the London system users can perform relevance feedback by moving keyframes around the screen.

All systems offer the user a way to visually inspect the result set as a list of keyframes ranked in the order of

similarity with the query (e.g. Fig. 2). The system from London ([12]) has an alternative view where the layout of keyframes on the screen visualises similarity of keyframes with the query. This is similar to our system, that visualises the similarity between keyframes (Fig. 3). Most systems use the temporal aspect of video by showing sequential keyframes of a shot within the ranked result list ([8, 10, 11, 13]). The London system ([12]) shows sequential keyframes in a separate window triggered by a user selecting a keyframe, similar to our series of consecutive keyframes on the bottom of Fig. 3. Most systems provide users with a way to inspect a single keyframe in a larger window ([8, 10–12]), and some let the user play the video ([8, 10, 11]).

We can conclude that there is a considerable overlap in functionalities. The querying and browsing interfaces show similarities across all systems. The main points that are specific to our system are the way of combining different retrieval mechanisms, and the use of LSI to facilitate concept search.

3 Methods

Prior to the study the subjects received three hours of training on the system. During the study, 21 groups of subjects (18 pairs and 3 individuals) searched the system for 12 topics per group. The data were analysed on the level of individual searches. A search is defined as the process of one subject group going through the three interactive stages of the system for one topic. After exclusion of searches that were not finished or contained too much missing data, 242 searches remained. To prevent sequential scanning of all shots in the collection, the time to complete one search was limited to 15 minutes.

Four types of data were gathered: average precision of a search, data about the interaction during a search, user estimation of the quality of a search, and the category of topics and queries.

3.1 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 result list [14]. AP lies between 0 and 1 and favours highly ranked relevant camera shots. Let $L^i = \{l^1, l^2, \ldots, l^i\}$ be a ranked version of the answer set A. At any given index i let $|R \cap L^i|$ be the number of relevant camera shots in the top i of L, where $|R|$ is the total number of relevant camera shots. Then AP is defined as:

$$AP = \frac{1}{|R|} \sum_{i=1}^{|A|} \frac{|R \cap L^i|}{i} \lambda(l^i) \quad (1)$$

where $\lambda(l^i) = 1$ if $l^i \in R$ and 0 otherwise. Note that AP is a quality measure for one search and not the mean quality of a group of searches. The iterative process of querying, browsing and ranking causes the AP of the result set to rise and fall during the search. Therefore, we recorded not only the AP at the end of the search but also the maximum AP during the search.

AP of each search was computed with a ground truth provided by TRECVID. Shots that were not in the ground truth were judged as being 'not relevant'. This is not always correct, since the ground truth contains only shots that were retrieved by the TRECVID participants. We do not consider this a problem since all searches in our study suffer from the same disadvantage.

3.2 Search data

In order to answer the first research question, logs of user interactions with the system were made containing the following data about each search:

1. Duration of the search.
2. Number of textual queries.
3. High-level concepts that were used.
4. Number of images selected.
5. Whether AND or OR search was used.
6. Whether exact (without LSI) or concept (with LSI) search was used.

These data were examined at two points in time: at the end of the search and at the point at which maximum average precision was reached.

3.3 User estimation

To answer the second research question, a questionnaire was developed to measure user estimation of the success of a search. Four questions were answered after each search:

1. Was it easy to get started on this search?
2. Was it easy to do the search on this topic?

Table 1: Summary of topics, categorised 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	01: aerial view of buildings and roads	09: the Mercedes logo
	06: helicopter in flight or on ground	25: the White House
	10: one or more tanks	07: Tomb of the Unknown Soldier
	13: flames	17: the Sphinx
	14: snow-covered mountains and sky	24: Pope John Paul II
	16: road(s) with lots of vehicles	04: Yassar Arafat
	18: a crowd in an urban environment	20: graphic of Dow Jones
	22: cup of coffee	15: Osama bin Laden
	23: cats	19: Mark Souder
Dynamic	05: airplane taking off	02: basketball passing down a hoop
	12: locomotive approaching you	03: view from behind catcher
	08: rocket taking off	while pitcher is throwing the ball
	11: person diving into water	

3. Do you expect that the results of this search contain a lot of non-relevant items (low precision)?
4. Are you satisfied with your search results?
In addition, subjects answered after each search:
5. Are you familiar with this topic?

All questions were answered on a 5-point scale (1 = not at all, 5 = extremely).

3.4 Categories of topic descriptions and textual queries

The 24 topics provided by TRECVID and the textual queries formulated by the subjects were categorised using a framework that was designed for a previous study [15]. The framework combines different methods (e.g. [16] and [17]) to categorise image descriptions into various levels and classes. For the present study we used only those distinctions that we considered relevant to the list of topics: 'general' against 'specific' and 'static' against 'dynamic'. Other distinctions, such as 'object' against 'scene', were not appropriate for the topic list since most topics contained descriptions of both topics and scenes. A summary of categorised topics is provided in Table 1.

4 Subjects

All subjects were 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 subjects were female, 37 male. Ages were between 20 and 40 (mean = 23.4).

To control in how far prior search experience might interfere with the effect of search actions on the results, we asked the subjects to fill in a questionnaire that contained 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. We did not find any evidence of a correlation between prior search experience and actions, nor between prior search experience and search results.

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 performing. All but three students indicated 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. The lack of influence of search experience can in part be explained by the fact that the system was different from search systems that the students were used to. 78% felt that the system was easy to use.

The subjects indicated a high familiarity with the topics (an exception was topic 19 'Find shots of congressman Mark Souder', with whom none of the participants was familiar). 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.

5 Results

5.1 Search data

The first research question was 'what search actions are performed by users and which actions lead to the best result?' In Table 2 descriptives of the six data types are presented that were recorded in the user logs. It shows that a search took approximately 8 minutes; a mean of 7.5 different textual queries were formulated during a search; a mean of 9 images were selected per search; high-level concepts were hardly used; the OR search was used more than the AND search; concept search (with LSI) was used in most cases. Time to finish topic, 'AND/OR search' and 'exact/concept search' did not affect the AP of the result. The remaining three variables are discussed below.

5.1.1 Query (re)formulation: In total, the subjects formulated 2141 textual queries. This brings the mean number of textual queries per search to more than seven.

Going back and forth between the different stages of the retrieval process, and reformulation of the query, is apparently an important part of user behaviour. This corresponds to the findings of Goodrum *et al.* [18], who examined image searching behaviour of users on the web. Query reformulation was one of the frequently occurring patterns of search tactics that was discovered. The number of queries did not affect the AP of the result.

5.1.2 High level concepts: The number of high-level concepts that was used in a search had a negative influence on the result. This is depicted in Fig. 4. The number of uses per high-level concept was too low to draw conclusions about the quality of individual concepts. We can conclude, however, that selection of more than one concept leads to low average precision. To give an indication of how the concepts were used by the subjects, Table 3 shows the frequency of use of the concepts and the mean AP of searches using the concepts. Only searches in which a single concept was used are included. Improving the precision of the concepts might lead to more use of the concepts and better results when concepts are combined.

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

		Max				End			
User Action	N	Min.	Max.	Mean	sd	Min.	Max.	Mean	sd
Time to finish topic (s)	242	0	852	345	195	6	899	477	203
No. of query (re)formulations	220	.	.	.	.	1	25	7.51	5.31
No. of high-level concepts used	240	0	5	0.50	0.84	0	17	0.59	1.39
No. of images selected	242	0	30	8.47	7.01	0	30	9.07	7.06
AND or OR search	240	AND:75 OR:165				AND:82 OR:158			
Exact or Concept search	240	Exact:69 Concept:166				Exact:62 Concept:176			

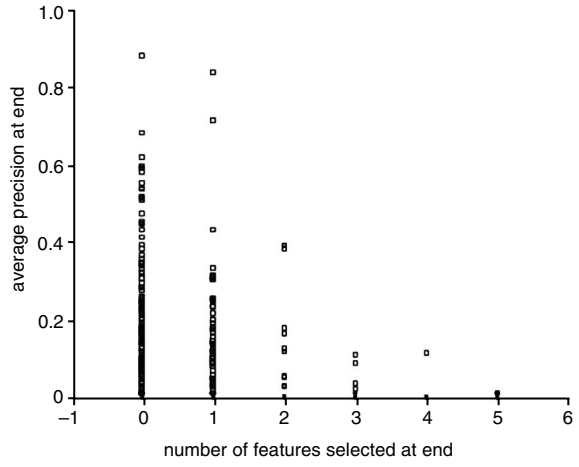

Fig. 4 *Scatter plot of number of selected concepts and AP at the end of the search*

One case with 17 concepts and AP of 0.027 is left out of the plot

Snoek *et al.* showed a great improvement in concept performance in [19].

5.1.3 Visual queries: The number of selected images was the most important variable to explain the result of a search (Pearson's correlation coefficient $r = 0.37$, $\alpha < 0.01$). This can be explained from 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 small; change in AP caused by the ranking step had a mean of 0.001 and a standard deviation of 0.032. The mean average precision at the end of a search was 0.16. The number of selected images was not correlated to the time to finish a topic, to the number of high-level concepts used, or to the type of search.

5.2 User prediction of search quality

5.2.1 User estimation: We collected opinions and expectations of users on each search. All questions measure an aspect of the user's estimation. For each question 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 four questions into one variable using principal component analysis (PCA). 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 each question on the first principal component. Pearson's correlation test showed a relationship between combined user estimation and actual measured average precision. ($r = 0.298$, $\alpha = 0.01$). This suggests that users are indeed able to estimate the success of their search.

5.2.2 Time between maximal AP and the end of the search: 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 Section 5.1, 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 (min = 0; max = 704, sd = 142). This means that students typically continued their search for 128 seconds (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. Not knowing when to stop searching is a general problem of category 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 ($r = -0.426$, $\alpha = 0.01$). 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 (min = 0; max = 0.48; sd = 0.043).

5.3 Topic and query category

5.3.1 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, which can be explained by the fact that our system treats the video data in terms of keyframes, which are still images. The differences were tested with

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

Table 3: High-level concepts: number of times a concept was used, mean average precision of searches using this concepts, and standard deviation of the average precision

Concept	N	Mean AP	sd	Concept	N	Mean AP	sd
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				

Table 5: Mean AP of topics types and ANOVA results

Mean AP	Static	Dynamic	Total	ANOVA results	SS	df	MS	F	Sig.
General	0.12	0.10	0.11	Between Groups	0.426	1	0.426	18.109	0.000
Specific	0.27	0.08	0.22	Within groups	5.648	240	0.024		
Total	0.19	0.10	0.16	Total	6.074	241			

Table 6: Query categories: absolute numbers at maximum AP and at the end of the search, and table percentages

Query type	From topic	Not from topic	Total
General	93 + 78(37%)	68 + 79(32%)	161 + 157(69%)
Specific	42 + 49(20%)	28 + 25(11%)	70 + 74(31%)
Total	135 + 127(57%)	96 + 104(43%)	231 + 231(100%)

an analysis of variance (shown only for specific/general topics). There is a strong correlation between topic-category and query-category. However, the correlation between topic-category and AP is valid regardless of the query-category used.

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

5.3.2 Query type: The length of textual queries varied from 1 to 22 words. To avoid domination of the analysis by long queries, we used only the first word of every query to determine the query type. During a search, users may formulate multiple textual queries. We analysed the last query before the moment of maximum AP, and the last query of the search (Table 6). 69% of the queries formulated by the subjects were 'general', 31% 'specific'; 93% were 'static', 7% 'dynamic'. Considering the low number of dynamic queries (only 16), we limit further analysis to the distinction between 'general' and 'specific' queries.

57% of the query words were copied directly from the topic descriptions. In the copied queries, the share of specific terms was higher than in queries that were not taken from a topic. An analysis of variance showed that 'specific' queries led to better results than 'general' queries ($F = 30.114$, $\alpha < 0.01$). This is still true for the 'general' topics: some participants used specific queries to search for general topics (e.g. query for Micheal Jordan, when looking for shots of basketball games), and that strategy worked very well. We did not find any evidence that other user actions were different for different topic categories.

6 Discussion

This study was concerned with the question how users search for news video in an interactive video retrieval system, and what factors influence the quality of their search results. The results showed two aspects of a user's search behaviour that positively affect the results: the number of selected images and the type of textual query.

The study has been carried out in one domain (broadcast news) using one retrieval system. Future research is needed to see whether the results can be extended to other domains and systems. Our expectation is that the broad domain of news will capture a lot of the difficulties in other domains. The specific structure of news videos - short stories about one topic - was not used by the retrieval system. A comparison of the user interface of the present system to user interfaces of other systems showed considerable overlap in functionalities. This strengthens our belief that the conclusions and recommendations that we present in this Section extend beyond this one system.

6.1 Textual queries

The contribution of the ranking step to the average precision was extremely small. From this we can conclude that text is a central feature in news video retrieval. This might change over time as performance of CBIR improves. The importance of text for video retrieval has not gone unnoticed in the TRECVID conferences and was pointed out by Hauptmann [20], amongst others. Eakins pointed out in [21] that users of image retrieval systems rate text entry interfaces higher than CBIR techniques such as QBE. This point should be taken into account when designing user interfaces of retrieval systems. Supporting text searches could, for example, be done by highlighting the words in the retrieved shot that match the user's query.

6.2 Topic type

'Specific' topics are better retrieved than 'general' topics. This is in accordance to the average TRECVID results [9]. In our study, the ranking step had a small but positive effect on general topics, while it had a small negative effect on specific topics. This suggests that a different strategy is optimal for different topic types: emphasis should be more on text for specific topics, while it can be on both text and low-level visual features for general topics. Yang *et al.* [3] also found that text is especially important for specific topics, while text and QBE are both of importance to generic topics. Letting the user adjust the weights of the two retrieval mechanisms, as is done by [10] and [11], is a good solution for expert users, but not for beginners as it requires the user to know about the strengths and weaknesses of the retrieval mechanisms. Future retrieval systems could benefit from a classification (either automatic or manual) of the topics, in order to adapt the retrieval strategy.

6.3 Browsing

The results show that from all recorded user actions, the number of selected images is the most important variable by far to explain the result. We conclude from this that the main contribution of content-based image retrieval to the retrieval process is visualisation of the dataset, which gives the user the opportunity to select manually relevant keyframes. The visualisation of the dataset 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 success quite well, but do not know when the optimal result is reached within a search. Effective visualisation of the dataset and improved facilities for browsing are therefore essential in future retrieval systems. In [22] an improved version is described of the current

similarity-based visualisation of our system (Fig. 3), that gives a better overview of the dataset. Owing to the recent nature of automatic retrieval systems, not much is known about the effectiveness of browsing interfaces for video. Van Houten *et al.* presented new ideas for a browsing interface in [23]. It would also be interesting to compare the results of an interactive video retrieval system to sequential scanning of shots in the dataset for a fixed amount of time.

6.4 Background knowledge

Prior experience with searching did not affect the quality of the search results. A possible effect could have been obscured by the three-hour training before the study, or by the fact that most subjects worked in pairs. However, Fang and Salvendy reported similar results in [24]. In their study, prior experience with search tools did not affect the success of searches on the web. Likewise, familiarity with the topic did not affect the quality of the search results. This seems to indicate that background knowledge of the searcher about the topic cannot be used adequately in the search process of current retrieval systems. Some attempts to include background knowledge into the process of multimedia retrieval have been made (see for example [25, 26]), but inclusion of background knowledge in interactive video retrieval systems is still at an early stage. We believe that text-based search could benefit from structured background knowledge in the form of ontologies or thesauri. This could, for example, be done by linking words in the query to concepts in an ontology, so that synonyms, related terms, broader and narrower terms can be found. In a similar fashion, we expect that searches using detection of high-level concepts could benefit from ontologies; by linking each detectable concept to a concept in the ontology, mutual relationships between the concepts can be exploited.

7 Acknowledgments

This paper was supported by the IOP Project 'interactive disclosure of Multimedia Information and Knowledge,' funded by the Dutch Ministry of Economic Affairs.

8 References

1 Hecht, A., O'Dwyer, A., Oomen, J., and Scharinger, F.: 'Birth: Building an interactive research and delivery network for television heritage', in Bearman, D. (Ed.): Proc. Int. Cultural Heritage Informatics Meeting, Berlin, 31 August – 1 September 2004
2 Christel, M., and Moraveji, N.: 'Finding the right shots: Assessing usability and performance of a digital video library interface'. Proc. ACM Multimedia, New York, USA, October 2004, pp. 732–739
3 Yang, M., Wildemuth, B.M., and Marchionini, G.: 'The relative effectiveness of concept-based versus content-based video retrieval'. Proc. ACM Multimedia, New York, USA, October 2004
4 Worring, M., Nguyen, G.P., Hollink, L., Gemert, J., and Koelma, D.C.: 'Accessing video archives using interactive search'. Proc. Int. Conf. on Multimedia and Expo, Taipei, Taiwan, June 2004
5 Hollink, L., Nguyen, G.P., Koelma, D.C., Schreiber, A.Th., and Worring, M.: 'User strategies in video retrieval: a case study'. Proc. Int. Conf. on Image and Video Processing, (Springer-Verlag Heidelberg), Dublin, Ireland, July 2004
6 Gauvain, J., Lamel, L., and Adda, G.: 'The limsi broadcast news transcription system', *Speech Commun.*, 2002, **37**, (1–2), pp. 89–108
7 Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R.: 'Indexing by latent semantic analysis', *J. Am. Soc. Inf. Sci.*, 1990, **41**, (6), pp. 391–407
8 Hauptmann, A., Baron, R.V., Chen, M.-Y., Christel, M., Duygulu, P., Huang, C., Jin, R., Lin, W.-H., Ng, T., Moraveji, N., Papernick, N., Snoek, C.G.M., Tzanetakis, G., Yang, J., Yan, R., and Wactlar, H.D.: 'Informedia at trecvid 2003: Analyzing and searching broadcast news video1'. TREC Video Retrieval Evaluation Online Proc., 2003
9 Smeaton, A.F., Kraaij, W., and Over, P.: 'TRECVID - an overview'. TREC Video Retrieval Evaluation Online Proc., 2003
10 Amir, A., Berg, M., Chang, S.-F., Hsu, W., Iyengar, G., Lin, C.-Y., Naphade, M., Natsev, A., Neti, C., Nock, H., Smith, J.R., Tseng, B., Wu, Y., and Zhang, D.: 'IBM research TRECVID-2003 video retrieval system'. TREC Video Retrieval Evaluation Online Proc., 2003
11 Browne, P., Czirjek, C., Gaughan, G., Gurrin, C., Jones, G.J.F., Lee, H., Marlow, S., McDonald, K., Murphy, N., O'Connor, N.E., O'Hare, N., Smeaton, A.F., and Ye, J.: 'Dublin City University video track experiments for TREC 2003'. TREC Video Retrieval Evaluation Online Proc., 2003
12 Heesch, D., Pickering, M.J., Ruger, S., and Yavlinsky, A.: 'Video retrieval using search and browsing with keyframes'. TREC Video Retrieval Evaluation Online Proc., 2003
13 Rautiainen, M., Penttil, J., Pietarila, P., Noponen, K., Hosio, M., Koskela, T., Mkel, S.-M., Peltola, J., Liu, J., Ojala, T., and Seppnen, T.: 'TRECVID experiments at Mediateam Oulu and VTT'. TREC Video Retrieval Evaluation Online Proc., 2003
14 National Institute of Standards and Technology (NIST): 'Guidelines for the TRECVID evaluation', http://www-nlpir.nist.gov/projects/tv2004/tv2004.html (last updated September 2004)
15 Hollink, L., Schreiber, A.Th., Wielinga, B., and Worring, M.: 'Classification of user image descriptions', *Int. J. Hum.-Comput. Stud.*, 2004, **61**, (5), pp. 601–626
16 Armitage, L.H., and Enser, P.G.B.: 'Analysis of user needs in image archives', *J. Inf. Sci.*, 1997, **23**, (4), pp. 287–299
17 Jorgensen, C.: 'Attributes of images in describing tasks', *Inf. Process. Manage.*, 1998, **34**, (2/3), pp. 161–174
18 Goodrum, A., Bejune, M.M., and Siochi, A.C.: 'A state transition analysis of image search patterns on the web', *Lect. Notes Comput. Sci.*, 2003, **2728**, pp. 281–290
19 Snoek, C.G.M., Worring, M., Geusebroek, J.M., Koelma, D.C., and Seinstra, F.J.: 'The mediamill TRECVID semantic video search engine'. TREC Video Retrieval Evaluation Online Proc., 2004
20 Hauptmann, A.G., and Christel, M.G.: 'Successful approaches in the TREC video retrieval evaluations'. Proc. ACM Multimedia, New York, USA, October 2004
21 Eakins, J.P., Briggs, P., and Burford, B.: 'Image retrieval interfaces: A user perspective', in Enser, P., Kompatsiaris, Y., and O'Connor, N.E. (Eds.): Proc. Third Int. Conf. on Image and Video Retrieval, Springer-Verlag Heidelberg, Dublin, Ireland, July 2004, pp. 628–637
22 Nguyen, G.P., and Worring, M.: 'Optimizing similarity based visualization in content based image retrieval'. Proc. IEEE ICME special session Novel Techniques for Browsing in Large Multimedia Collections, Taipei, Taiwan, 2004
23 van Houten, Y., Schuurman, J.G., and Verhagen, P.: 'Video content foraging'. Proc. Third Int. Conf. on Image and Video Retrieval, Dublin, Ireland, July 2004, pp. 15–23
24 Fang, X., and Salvendy, G.: 'Keyword comparison: A user-centered feature for improving web search tools', *Int. J. Hum.-Comput. Stud.*, 2000, **52**, (5), pp. 915–930
25 Jaimes, A., Tseng, B.L., and Smith, J.R.: 'Modal keywords, ontologies, and reasoning for video understanding', *Lect. Notes Comput. Sci.*, 2003, **2728**, pp. 248–259
26 Hyvonen, E., Saarela, S., Viljanen, K., Mkel, E., Valo, A., Salminen, M., Kettula, S., and Junnila, M.: 'A cultural community portal for publishing museum collections on the semantic web'. Proc. 16th Eur. Conf. on Artificial Intelligence (ECAI), 2004