Int J Multimed Info Retr (2012) 1:17–29
DOI 10.1007/s13735-012-0007-3

INVITED PAPER

Bridging the gap between expert and novice users for video search

Martin Halvey · Joemon M. Jose

Received: 14 January 2012 / Accepted: 1 February 2012 / Published online: 7 March 2012

Abstract Contemporary video retrieval systems are wanting in terms of helping users find appropriate videos. There are a number of reasons for this, including a lack of appropriate representations for video and the semantic gap. These problems are amplified by the fact that the importance of expertise for effective video search is not well understood. In an attempt to garner a greater understanding of the impact of expertise on video search, a user evaluation that is designed to investigate the role of expertise in video search was conducted. In our evaluation participants were given a number of video search tasks and were asked to find relevant videos using two different interfaces: the first interface required users to use background knowledge to find relevant videos and the second interface allowed users to use video search tools to complete the task. Three groups of users with varying search expertise carried out these video search tasks, with the objective that the behaviour and success of the different user groups could be examined. It was discovered that the behaviour of novice users begins to emulate that of the expert users as the novice users gain more expertise. However, it was also found that the perceptions of novice users, even with additional background knowledge, of the tools, collection and performance do not always match that of the expert users.

Keywords Video · Search · Navigation · Query formulation · Domain knowledge · Expertise · Expert

M. Halvey (✉) · J. M. Jose
School of Computing Science, University of Glasgow,
Sir Alwyn Williams Building, Glasgow, Scotland G12 8QQ, UK
e-mail: martin.halvey@glasgow.ac.uk

1 Introduction

There is a growing necessity for effective tools and techniques to assist users in the difficult task of searching for video. Current cutting edge video retrieval systems rely on various techniques to bridge the semantic gap, including using annotations provided by users, methods that use the low level features available in the videos or an existing representation of concepts associated with the retrieval task. None of these methods are sufficient to overcome the problems associated with video search (see Sect. 2 for a full discussion). These problems are magnified by the lack of understanding of user behaviour while searching for video clips. In order to address this issue, in this paper we introduce an experimental study to investigate the differences between expert and novice users when searching for video. Understanding the differences between the behaviour of these two user types could have a number of positive outcomes that will influence the development of tools and techniques for video search. e.g. improved search interfaces, improved query methodologies, a more diverse set of tools that can be used by all user types for search, etc., all of these close the gap between novice and expert users. Additionally, further understanding of user behaviour could influence the assessment of tools and techniques, which in turn could lead to more realistic and robust evaluations of these tools and techniques. These potential beneficial outcomes and the success of our evaluation are based on the assumption that more expert users should be more successful in finding relevant video than more novice users. For our evaluations we use the TRECVID collection, tasks and definition of expertise, and for a number of years in TRECVID expert users have outperformed novice users. This presents us with a unique opportunity to compare different user types (in terms of expertise) performance and approach for video retrieval and as such answer

some key questions regarding the role of expertise in video retrieval.

On the whole our objective is to provide answers to a number of questions regarding novice and expert users for video search. To begin with we wish to discriminate between the actions of novice and expert users. By analysing the ways in which both sets of users carry out video search tasks, we hope to identify the good practices that lead to successful searches by expert users and the actions that lead to unsuccessful searches by novice users. There are a number of ways that multimedia search behaviour may be influenced by expertise. First, background knowledge could affect the quality, quantity and type of queries used. This topic has been investigated extensively for the Web through log file analysis [1], in electronic databases [2] and briefly but not to a great extent for video search [3,4]. However, searching for multimedia provides a larger number of methods to query search systems when compared with other search paradigms (see the Sect. 2 for a full discussion). Second, background knowledge could affect the videos that users select to click on. In most retrieval systems video shots or entire videos are normally represented by one keyframe; a keyframe is an image that represents an entire video shot. Due to the volume of information that may be contained in even short video clips it may be the case that these selected image keyframes may not be representative of the video content. However, based on these keyframes users must decide about whether to view a video or not. Models of language comprehension have shown the importance of background knowledge when making judgements within text [5]. Indeed in the online domain, links followed during navigation have been investigated thoroughly [6]. However, in the visual domain where an image represents a retrieved video, this issue of background knowledge has not been as thoroughly investigated. Finally, with respect to the decision to follow a particular search direction, domain knowledge may affect the decision of a searcher to leave a particular search path, which in the case of video search would be ceasing to watch a particular video or set of video clips.

A second goal of this research is to examine which particular characteristic of being an expert user results in more successful searches. Is it greater knowledge of tools and the search system? Or is it indeed greater domain knowledge? The implications of this discovery could affect the design of video retrieval tools to help users. These tools should aspire to close the gap between the particular aspects of user searches that make the searches representative of a novice or expert.

A final objective is to examine if increased domain and/or tool knowledge can lead to improved novice performance and also in addition if it is possible for a novice user to perform as well as an expert user. Once again the implications of this finding could be beneficial as a guide for designing video search tools. Moreover, it is hoped that as well as influencing what types of tools will aid user search, the findings of this paper will aid the design and evaluation of multimedia search tools for a wide variety of users.

2 Related work

In recent years, as a result of the improving capabilities and the falling costs of many hardware systems, there are greater than ever resources to store and manipulate videos in a digital format. Additionally, with greater than ever broadband capabilities it is now practical to watch videos almost anywhere as simply as text-based pages. People can now create their own personal digital libraries from multimedia produced through digital cameras, mobile phones and camcorders etc., and use a variety of systems to place this material on the web, as well as gather these materials in their own personal multimedia collections. This has resulted in the emergence of numerous online video search and sharing sites, among them YouTube[1] and Blinkx.[2] Despite this rapid growth in the availability and the ubiquity of video, the systems that currently exist to organise and retrieve these videos are not sufficient to deal with such large and rapidly increasing volumes of video. In an effort to assist the development of tools and techniques to support video search, the TRECVID workshops provide a common test collection to enable large-scale evaluation of research approaches. In terms of interactive video search TRECVID participants have had a relatively great deal of success. However, a lot of this success is for expert users who form the upper bound for model users of interactive video search systems [4]. Also in recent years some of the video search systems have concentrated more on users categorising search results rather than actually aiding users in searching for video [7]. In order to examine the reasons behind the success of expert users and the difficulties involved with interactive video search, we will discuss interactive video search in more detail.

2.1 Interactive video retrieval

Interactive video retrieval refers to the process of users formulating and carrying out video searches and subsequently reformulating queries and getting new search results based on the previously retrieved results. As video is extremely rich content there are a number of different ways that users can query video retrieval systems. The use of the low-level features that are available in images and videos, such as colour, texture and shape to retrieve results, is one common approach. This approach is often used for query by example, where users provide sample images or video clips as examples to retrieve similar images or video clips. While

[1] http://www.youtube.com.

[2] http://www.blinkx.com.

this approach seems reasonable it also presents a number of problems. It requires a representation and extraction of all of the features required from all of the videos, presenting issues of efficiency. Also the difference between the low-level data representation of videos and the higher level concepts users associate with video, commonly known as the semantic gap [8], provides difficulties. In an attempt to bridge this semantic gap, a great deal of interest in the multimedia search community has been invested in search by concept. The idea is that semantic concepts such as "vehicle" or "person" can be used to aid retrieval; an example of this is the large-scale ontology for multimedia (LSCOM) [9]. However, query by concept also has a number of issues that hinder its use; it requires a large number of concepts to be represented and to date has not been deployed on a large-scale for general usage.

Query by text is the most popular method of searching for video. It is used in many large-scale video retrieval systems such as YouTube, Blinkx, etc., and is also the most popular query method at TRECVID [4]. Query by text is simple and users are familiar with this paradigm from text-based searches. In addition to this, query by text does not require a representation of concepts or features associated with a video, but does require that meaningful textual descriptions of the video and its content are available. Textual descriptions in some cases may be extracted from closed captions or through automatic speech recognition; however, a study of a number of state-of-the-art video retrieval systems [10] concludes that the availability of these additional resources varies for different systems. Where these resources are available, they may not always be reliable due to limitations in automatic speech recognition or language differences for example. More recent state-of-the-art online systems rely on using annotations provided by users to provide descriptions of videos. However, as was stated previously, this further complicates the retrieval process, either because of misconception surrounding annotations [11] or users' reluctance to provide annotations [12].

While all of these methods outlined above have problems, they have been used together in a number of systems, such as Informedia [13] and MediaMill [14]. These systems have been amongst the most successful systems at recent TRECVID interactive search evaluations. However, these top results are for "expert" users, who, as discussed earlier, establish an idealistic performance threshold for users [4]. Also, a combination of these approaches requires a vast amount of metadata to be extracted and stored for each individual video clip. As we have previously outlined, there are a number of different ways in which a user can query a video retrieval system; these include query by text, query by example and query by concept. Each of these approaches have had limited success on video retrieval, and to date none of these approaches has provided an adequate solution to providing the tools to facilitate video search [3]. Thus, in an attempt to overcome these limitations some innovative interfaces have been proposed for providing different types of interaction for multimedia search.

2.2 Interactive multimedia retrieval interfaces

In this sub-section we outline some innovative and interesting interfaces for interactive multimedia retrieval; we first discuss some image retrieval interfaces before discussing video retrieval. Image retrieval and video retrieval are similar in that in both cases there is a combination of visual and often textual information; however, there are a number of important differences that make video search much more difficult than image search. The first main difference is the multimodal nature of video, encompassing images, text, audio and a temporal factor, etc. While text and visual features may be used to aid or hamper image search, these are only two of the many modalities involved in video search. Second, video is a much more interactive medium in comparison with still images. Interactive video retrieval systems have to make an additional effort to aid the user in deciding whether the selected videos are relevant or not for their tasks, whereas for image retrieval systems the user can easily and quickly discern relevant and irrelevant results. The result of this is that interaction and usage information from interactive video retrieval systems are far noisier than the usage information on image retrieval systems. For instance, on average, 75% of the user results that a user interacted with on the image retrieval system developed by Craswell & Szummer were relevant [15], whereas only 7–9% of search results that the user interacted with were relevant for a similar interactive video retrieval system [16]. As a result of this, the goals of many interactive video retrieval systems are to lower the effort for the user to explore the complex information space and also to assist the user in deciding if a result is relevant to their information need. Despite these differences it is still important to consider the myriad of different image retrieval systems available while considering video retrieval, as both image and video retrieval systems attempt to bridge the semantic gap.

PicturePiper [17] provides a mechanism for allowing users access to images on the web related to a topic of interest. This system was developed to demonstrate a reconfigurable pipeline architecture that is ideally suited for applications in which a user is interactively managing a stream of data. PicturePiper also contains a workspace for displaying search results; the distance between groups of images in the workspace illustrates the distance between the centroids of the groups of images as calculated using low-level features. EGO [18] is a tool for the management of image collections; like PicturePiper EGO has a workspace, this workspace is augmented with a recommendation system. By providing these facilities, different types of requirements

are catered for, enabling the user to both search and organise results effectively. The workspace serves as an organisational ground for the user to construct groupings of images. The recommendation system in EGO observes the user's actions, which enables EGO to adapt to their information requirements and to make suggestions of potentially relevant images based on a selected group of images. The MediaGLOW system [19] presents an interactive workspace that allows users to organise photographs. Users can group photographs into stacks in the workspace; these stacks are then used to create neighbourhoods of similar photographs automatically. CueFlik [20] is a Web-based image search application that allows users to create their own rules for ranking images based on their visual characteristics. Users can then re-rank possible search results according to these rules. In user evaluations it was found that users can quickly create effective rules for a number of diverse concepts. Campbell presents a novel image search and browsing system in the Ostensive Browser [21]. The main component of the interface is a workspace with objects on it and links between those objects. The user begins browsing at a starting image, around which candidate next images are displayed. A user selects a candidate which becomes the centre of focus; the next possible candidate images related to the current image are displayed. Browsing continues in this fashion. Candidate images for browsing are determined by an ostensive model, which encompasses a temporal profile of uncertainty. This is accomplished by the application of a particular class of discount function with respect to the age of the evidence; thus more recent user interactions are more relevant for determining next steps in comparison with older interactions.

With respect to video retrieval the Fork Browser [14] embeds multiple search methods into a single interface for browsing. The multiple search methods are presented to the user in the form of threads. Each thread is a ranked list of shots based on one of the search methods implemented in the interface. The threads are visualised in the shape of a fork. The shot at the top of the stem of the fork is the video that the user is currently viewing, with the tines representing the different threads. The Extreme Browser [22] aims to maximise the human capability for judging visual material quickly, while at the same time applying active learning techniques using user selected videos. Videos are presented to the user via a method called rapid serial visual presentation which allows the user to make fast judgements about high numbers of videos. The feedback from the user is used in an active learning loop, which is used to rank the remaining results that the user will review. The FacetBrowser [23] is a video search interface that supports the creation of multiple search "facets", to aid users carrying out complex video search tasks involving multiple concepts. Each facet represents a different aspect of the video search task: an assumption of this work is that search facets are best represented by sub-searches. These facets can be organised into stories by users, facilitating the creation of sequences of related searches and material which together can be used to satisfy a work task. The interface allows more than one search to be executed and viewed simultaneously, and importantly, allows material to be reorganised between the facets, acknowledging the inter-relatedness which can often occur between search facets. ViGOR [24] is a video retrieval system that allows users to group videos to facilitate video retrieval tasks. In this way users are able to visualise and conceptualise many aspects of their search tasks and carry out a localised search to solve a more global search problem.

This section has presented an introduction to problems associated with video search and some systems that attempt to overcome this problem; a more complete introduction to video retrieval interfaces can be found in Schoffman et al. [25].

2.3 Expert search

Comparisons between expert and novice users have been carried out in a number of fields. In library studies Hsieh-Yee [26] found that increased search expertise results in different behaviours for different users. Librarians used synonyms, thesaurus terms and more combinations of search terms when searching within a library database, in comparison with novice users. Such search behaviour increased with subject or domain knowledge, but this effect was only found for librarians and not for novice users. Studies have also been conducted to investigate the effect of search expertise and domain knowledge on the early web [27]. The results of this study found, similarly to [26], that both expertise and domain knowledge can lead to improvements in query formation and selection. More recently, Lazonder et al. [28] found that general experience searching the Web can result in faster and more successful location of websites. Bhavnani [29] looked at the effect of expertise on the web in more detail. This study investigated the behaviour of a number of users with domain expertise (healthcare and online shopping) and also general Web experience. It was found that all of the participants were more proficient within their own area of expertise than with other topics, mainly achieved by utilising domain knowledge, e.g. using specialist websites. More recently, Duggan and Payne [30] investigated the importance of background knowledge for carrying out effective web searches. They found that knowledge of a topic predicted search performance on that topic, including for questions for which participants did not already know the answer. It was also found that greater topic knowledge resulted in less time being spent on a webpage, faster decisions being made on when to abandon a search and shorter queries being submitted to the search system.

However, the importance of background knowledge for effective video search has not been studied extensively and thus is not well understood. Christel and Conescu [3] investigated the use of specific search features for novice users in TRECVID 2005. It was found that the availability of more shot and image-based search tools resulted in these visual tools being used more widely in comparison with text. It was also discovered that suppressing previously viewed shots had a negative effect on user performance. Following on from this study Christel [4] investigated how more visual search tools could be used to close the gap between the performances of novice users in comparison with users who have considerable domain knowledge. In fact it was found that novice users could outperform expert users when the topics fell outside the area of the expert's domain knowledge.

3 Experimental methodology

3.1 Collection and tasks

The TRECVID 2007 video collection and tasks were utilised for this evaluation. There are a number of reasons for using this collection. First, the TRECVID collection is a large and well-known video collection. In addition to TRECVID topics there exists a well-defined definition for novice users and expert users. A modified version of the TRECVID user definition was used for this evaluation; this will be outlined in the next section. Second, the TRECVID collection has a number of diverse tasks for which the relevant video shots in the collection have been marked. Finally, previous TRECVID collections have also been utilised to provide an analysis of interface design properties for video search for expert users and novice users [3,4]. In 2007 the TRECVID collection contained 18,142 shots (over 100 h) of Dutch magazine television. For the TRECVID 2007 interactive search evaluations there were a total of 24 tasks. For this evaluation the number of tasks that the users carry out is limited to 8. Reducing the number of tasks to eight allowed more evaluations to be conducted, as 24 individual search tasks did not have to be carried out for each participant, while at the same time still providing a robust comparison between users. In order to examine user interactions on different types of tasks the eight chosen were the tasks that had the highest number of shots marked as being relevant during TRECVID runs (in comparison with the 16 other tasks). This choice was made for a couple of reasons. First, some of the participants were novices we wanted to ensure that there were a sufficient number of possibly relevant video shots, so that we could compare the success and practises of different types of users. Also we did not want participants to become frustrated at the difficulty in finding shots. Second it could be postulated that as there are more relevant video shots that there are more ways of finding these shots and a wider variety of shots to find, as there is more variance in the shots than in a smaller selection. The eight tasks used for this evaluation are (for a more simple presentation we have numbered the tasks 1–8 for our evaluation; for each task we presented the users with a topic number, topic and example images, see Fig. 1):

- Task 1: Find shots of a person walking or riding a bicycle (Topic Number 0199, 1,175 relevant shots)
- Task 2: Find shots of a woman talking toward the camera in an interview—no other people visible (Topic Number 0213, 400 relevant shots)
- Task 3: Find shots of one or more people playing musical instruments such as drums, guitar, flute, keyboard, piano, etc. (Topic Number 0218, 376 relevant shots)
- Task 4:Find shots with hills or mountains visible (Topic Number 0206, 343 relevant shots)
- Task 5: Find shots with 3 or more people sitting at a table (Topic Number 0209, 332 relevant shots)
- Task 6: Find shots of waterfront with water and buildings (Topic Number 0207, 265 relevant shots)
- Task 7: Find shots of a very large crowd of people (fills more than half of field of view) (Topic Number 0214, 264 relevant shots)
- Task 8: Find gray-scale shots of a street with one or more buildings and one or more people (Topic Number 0220, 210 relevant shots)

3.2 Participants

16 participants participated in the evaluation: eight were classified as novice and eight were classified as experts prior to the evaluation. There are a number of possible definitions for an expert in video search. As was stated earlier for the purposes of this evaluation and also for simplicity a modified version of the TRECVID definition of an expert was used. All of the experts had to meet each of the three following criteria:

1. The expert has been active within the multimedia IR community for at least a year, having a better sense of the accuracy of the various automated video processing techniques
2. The expert has used similar video retrieval systems prior to the timed runs on the TRECVID data and therefore is more familiar with the system operations than study participants who first see it during the evaluation
3. The expert knows about TRECVID evaluation, e.g. the emphasis on shot-based retrieval and the use of mean average precision as a key metric, etc.

While we are aware that this definition may be somewhat narrow, it is a well-defined definition which has been widely

Topic Number 0206

Find shots with hills or mountains visible

Image examples:

Fig. 1 Example task as presented to experiment participants

used for the evaluation of state-of-the-art video systems at the TRECVID workshop [31]. If the evaluation were on a different collection, e.g., YouTube or television archives in an industrial setting, then perhaps a different definition of an expert may be more appropriate.

The expert group consisted of eight males, all of whom were members of our research group. The novice group consisted of five males and three females; the novice users were mainly students in the Computing Science department but also included people from other professions, e.g., a musician and a marketing executive. The average age of the members of the expert group was 27 with the average of the members of the novice group being 26. All of the expert users had extensive experience of dealing with multimedia search tools in both work and leisure. Similarly, the novices also had experience of dealing with multimedia search tools; many of the users in both groups cited Flickr[3] and YouTube as being services that they use very often.

3.3 Video retrieval system

With the aim of evaluating the effect of domain and tool knowledge on user search behaviour and performance, two interfaces for video search were utilised for the evaluation. The first interface is called the search interface (SI). The SI consists of basic tools for video search, which include query by text and query by example. This interface puts the emphasis on the user to use their own knowledge to perform search tasks; this permits an assessment of the impact of collection knowledge on user actions and interactions. The second interface is the ViGOR system [24], which was outlined in the related work section. ViGOR allows users to create groups of video shots to assist with completing their search and also allows users to use these groups as a starting point for further searching to achieve their search goal. Previous evaluations have shown that ViGOR results in increased user performance and increased user satisfaction [24]. The purpose of ViGOR in this evaluation is to provide a system which can supply information about how different users conceptualise different searches. Also as ViGOR is a slightly more complex video search system, the user interactions can potentially provide further information about how different user types utilise more complicated tools when they are available. In addition, the different focus of both interfaces and indeed the use of two interfaces provide a simulation of a typical video retrieval experimental setup, where more often than not, more than one interface or system is being investigated in order to examine an effect.

In brief, user interaction with the SI should supply information about how users make use of domain knowledge and user interactions with ViGOR should provide us with data about users' use of more complicated tools and how users structure their searches.

The SI (see Fig. 2 and the search and results panel on the left of Fig. 3) consists of a search panel (A) and a results display area (B). The search panel is where users carry out their searches. Users enter a text-based query in the search panel to begin their search (a). The users can add videos from the results as examples or enter text to reformulate their queries to continue the search process (b). The result panel is where users can view the search results (c). Users can drag shots from this panel and add them as example shots to reformulate their query (b). Users can also drag shots from this panel and add them as relevant images for the task (h). In all panels additional information about each video shot can be retrieved. Hovering the mouse cursor over a video keyframe will result in that keyframe being highlighted, along with neighbouring keyframes and any text associated with the highlighted keyframe (henceforth referred to as tooltip). If a user clicks on the play button a popup panel appears to play the highlighted

[3] http://www.flickr.com.

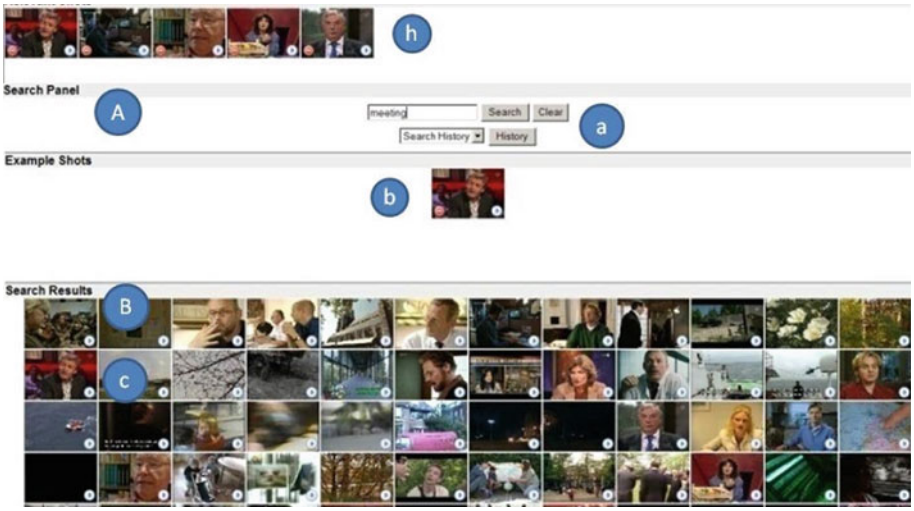

Fig. 2 Screen shot of the search interface

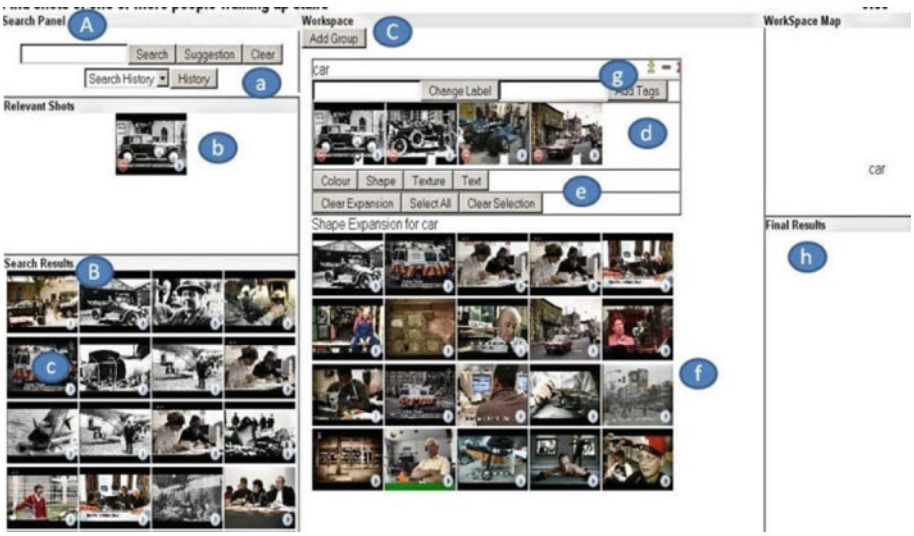

Fig. 3 Screen shot of ViGOR. The search interface system consists of the search panel and results panel that are on the *top left* of ViGOR

video shot. As a video is playing it is possible to view the current keyframe for that shot, any text associated with that keyframe and the neighbouring keyframes. Users can play, pause, stop and navigate through the video as they can on a normal media player.

ViGOR (see Fig. 3) comprises of a search panel (A), results display area (B) and workspace (C). These facilities enable the user to both search and organise results effectively. The search panel (A) and the results panel (B) (both on the left-hand side of the interface) are where users can view the formulated queries and search results, these search facilities can also be found on the SI. The main component of ViGOR which differentiates it from the SI is the provision of a workspace (C). The workspace serves as an organisational ground where the user can construct groupings of images. Groups can be created by clicking on the create group button; the users then add a textual label for the group (g). Users can potentially add an infinite number of annotations to the group, but each group must have at least one annotation. Drag-and-drop techniques allow the user to drag videos into a group (d) or reposition the group on the workspace. It should be noted that any video can belong to multiple groups simultaneously. Each group can be used as a starting point for further search queries. Users can select particular videos and can choose to view similar videos (f) based on one or all of a set of feature categories [colour, texture, shape or text, respectively (e)]. The workspace is designed as a potentially infinite space to accommodate a large number of groups. ViGOR also contains the same play and tooltip functionality as the SI.

3.4 Experimental design

A 2-searcher-by-2-task Latin Square design was adopted for this evaluation. Each participant carried out two tasks using the SI, and two tasks using ViGOR. In order to avoid any order effect associated with the tasks or with the systems

the order of system usage was varied as was the order of the tasks. With the goal of determining the effect of domain and tool knowledge, each novice user returned to carry out four different tasks on the day following their original evaluation; this produced a third group of distinct users which will henceforth be referred to as Novice+. Using this experimental model it is possible to evaluate the potential effect of additional knowledge of the collection and/or the tools provided on user search behaviour and interactions. Additionally, the ground truth provided in the TRECVID 2007 collection allowed us to conduct analyses that may not have been possible with other collections.

Each participant was given 5 min training on each system and every participant was allowed to carry out a training task using each interface. These training tasks were also tasks from TRECVID 2007; these tasks were judged to be appropriate as they also had relatively high numbers of relevant videos. Each actual task had a 15 min time limit. Every participant had their interaction with the system logged, the videos they marked as relevant were stored and they also filled out a number of questionnaires at different stages of the experiment.

4 Results

4.1 Task performance

As the TRECVID 2007 collection was used for this evaluation it is possible able to calculate precision and recall values for all of the tasks. Table 1 presents some of the results for all three types of users and both systems, for each user type Table 1 shows the number of documents found, the number of relevant documents found and the mean average precision (MAP). MAP is the average for 11 fixed precision values of the precision/recall metric and is normally used for a simple and convenient system performance comparison. As MAP is calculated using measures that relate to both recall and precision it gives an overall measure of the performance of the system.

Based on the results presented in Table 1 it is apparent that the expert users outperform the novice and novice+ users in all reported aspects of task performance. The expert users find more video shots that they believe to be relevant, find more relevant video shots and result sets found by the expert users also have the highest MAP score, which is an indication of the best overall performance. On the other hand, for some of the tasks it was found that novice and novice+ users had higher precision for the retrieved video shots; however, this gives a false impression in relation to task performance as this was normally for low numbers of videos.

Another interesting finding is that the performance of the novice users appears to improve greatly when the users return to perform retrieval tasks as a novice+ user. The figures in Table 1 indicate that the performance of the users in relation to all reported aspects of task performance improves. In order to investigate if the increase in performance was significant a multi-way analysis of variance (ANOVA) with system, user type and task as the factors was performed. It was found that for MAP, videos found and relevant videos found that task was a significant factor. This result is not surprising as the tasks are very different and users indicated that they found the difficulty of tasks to be variable. However, it was also discovered that the difference in the number of videos retrieved was statistically significant for the user type. In order to expand on these results the different user types were paired to see if there was any effect; the difference in the number of videos retrieved was statistically significant between novice and expert users ($F = 4.622$, $p = 0.0359$), but not between novice+ and expert users and novice+ and novice users, although there is a definite trend between these user types that performance increases with expertise. These results show that the biggest difference in performance is between expert and novice users and that it appears that the novice+ users with extra exposure to the system and collection can perform the video search tasks almost as well as the expert users who have a more comprehensive understanding of the system and collection. Despite the fact that the differences in relevant videos found and MAP are not statistically significant, the same tendency can be seen for all the performance measures, the expert users have the best overall performance, however, with some additional knowledge the novice+ users can perform almost as well as the expert users.

Table 1 Task performance statistics for expert, novice and novice+ users for the SI and ViGOR

User type	Videos found	Relevant videos found	MAP
Overall			
Expert	25.687	15.187	0.033
Novice	16.782	10.969	0.025
Novice+	23.594	13.187	0.028
Search interface			
Expert	26.562	17.5	0.039
Novice	19.75	13.562	0.032
Novice+	25.312	15.187	0.033
Grouping interface			
Expert	24.812	12.875	0.027
Novice	13.187	7.812	0.017
Novice+	21.875	11.187	0.023

4.2 User behaviour

In order to understand the improved performance of novice+ and expert users in comparison with novice users an

Table 2 Query behaviour for expert, novice and novice+ users for the SI and ViGOR

Search action	Expert	Novice	Novice+
Query length	1.329	1.387	1.289
With image example	231	350	236
No examples	2.095	3.003	2.186
Group image	283	200	163
No. examples	1.849	2.039	2

analysis of the approach and behaviour of different users while carrying out their search tasks was conducted

4.2.1 Query formulation

In order to investigate if there was any difference in user behaviour for query formulation a number of aspects of user queries were investigated, including query length, the number of times images were used in a query, the average number of example images used, the number of times group of images were used to launch a search and the average number of examples used. This analysis (see Table 2) revealed that expert users have the shortest textual queries typically, but subsequent to having used the search tools the novice+ users shorten their queries. Also, expert users are less likely to add an example image to their query and when expert users do add examples images they add less example images in comparison with novice users. Once again it can be seen that after using the system novice+ users query creation behaviour changes; the novice+ users begin to carry out less searches with a query examples and also with less examples than when they were novice users, in a way their behaviour is becoming more like that of an expert user.

When it comes to using the additional search functionality available in the grouping interface, i.e. using groups to launch a query using a low-level feature, it was found that experts use these features more often than any other user group. What is even more surprising is that the novice+ users actually use this feature less often than when they were members of the novice user group. This might be part of an overall trend for the novice+ users, given that the novice+ users also carry out fewer searches with image examples as well. Nevertheless, it should be noted that despite the fact that novice+ users use the additional search functionality available in the groups less than when novice users, they do use less query images than novice users; thus their behaviour mirrors expert users in this respect.

An additional fascinating trend appears when the additional search features available for groups in ViGOR are examined, in particular which features are used most often by users for the different tasks. As has been seen thus far, the expert users are the most successful users in our evaluation; given this fact and their background knowledge, the assumption could be made that the expert users would be most likely to be able to determine the most useful additional search feature available for each task. Table 3 shows the search feature used most often for each user type for each task. It can be seen in Table 3 that the expert users and novice users used the same feature most often on two occasions; in contrast with this, expert and novice+ users use the same feature most often for five of the eight tasks. Novice and novice+ users are in agreement three out of eight times. Overall, all user groups utilise the same feature most often for only one of the tasks; this is task eight where users must retrieve gray-scale images, so the fact that they pick the same feature (colour) in this instance is not surprising. The results in Table 3 once again demonstrate a change in the behaviour of novice+ users, i.e. when the novice users return and are in the novice+ group where they have more background knowledge, their behaviour becomes more like that of an expert user.

Table 3 Predominant query type for expert, novice and novice+ users by task for ViGOR

Task	Expert	Novice	Novice+
1	Colour	Colour	Shape
2	Colour	Shape	Colour
3	Colour	Shape	Shape
4	Colour	Texture	Colour
5	Colour	Shape	Shape
6	Colour	Texture	Colour
7	Texture	Shape	Texture
8	Colour	Colour	Colour

4.2.2 Video navigation usage

As has been demonstrated thus far, the search behaviour of the novice users changes as they gain more background knowledge of the collection and the search tools. The analysis of user behaviour continues by looking at the navigation behaviour of different user types and also their use of the available search tools. Table 4 shows the average number of times that each user type plays a video, navigates through a video (using the next or previous function) to view a neighbouring shot and uses the tooltip functionality or creates a video group.

Once again it can be seen, based on the figures in Table 4 that the behaviour of the novice users changes when they return as a novice+ user to become closer to the behaviour of an expert user. The expert users use the navigational tools more often than the novice users; they play and navigate

Table 4 Use of navigation and search tools for expert, novice and novice+ users for the SI and ViGOR

User	Play	Prev	Next	Tooltip	Group
Overall					
Expert	32.656	25.25	54.187	80.906	2.75
Novice	29	19.812	46.968	74.843	1.687
Novice+	28.687	44.34	51.625	78.343	1.875
Search system					
Expert	34	25.875	59.875	79	N/A
Novice	30.875	20.937	55.062	75.812	N/A
Novice+	30	54.5	58.75	78.062	N/A
ViGOR					
Expert	31.312	24.625	48.5	82.812	2.75
Novice	27.125	18.687	38.875	73.875	1.687
Novice+	27.375	34.187	44.5	78.625	1.875

through more videos and also use the tooltip function more often (to gain information about the video shots). The expert users also create the most video groups; in addition, the expert users use these as points to launch searches more often than novice users (see Table 2). The novice users perform all of these actions the least often of all of the user types, with the novice+ user's behaviour changing to be closer to that of the expert users with additional knowledge of the tools and the collection.

This change in the behaviour of the novice users with respect to the search system could indicate that the users are more aware of the contents of the collection. The SI requires users to utilise their knowledge of the video collection to devise effective search queries to search the collection. The change in behaviour of the novice users with respect to ViGOR is an indication that the novice+ users are more aware of the benefits of tools available and are also confident enough to use the tools provided. ViGOR includes more tools which allow users to express themselves and to delve deeper into the collection, while at the same time making it easier for users to query the collection. The results in these last two sections indicate that when the novice users return for a second session as novice+ users they become more aware of the collection and more confident with the tools, with increased knowledge of both. The change in behaviour and difference in performance for both of ViGOR and SI system indicate that perhaps the novice+ users' change in performance and behaviour cannot be accredited to a simple gain in knowledge about the tools or about the collection, but perhaps in both. These findings potentially have a number of implications for the design and evaluation of multimedia information retrieval systems for all user types. In order to explore these findings in more detail we turn to the user questionnaires.

Table 5 User perceptions of tasks for each user type, ignoring system (Higher = Better)

Differential	Expert	Novice	Novice+
Clear	4.625	**4.843**	4.812
Easy	3.187	3.469	**3.844**
Simple	3.812	3.812	**4.156**
Familiar*	**4.156**	3.281	3.75
Relaxing	3.25	3.406	**3.656**
Interesting	3.281	3.351	**3.469**
Restful*	2.875	3.094	**3.5**
Easy*	3.187	3.312	**3.906**

Statistically significant differences are marked with $*p < 0.05$

4.3 User feedback

4.3.1 Task and search perception

As part of the post task questionnaire the users were asked about the tasks and the search that they had carried out. The following semantic differentials were used on a 5-point scale:

- The task you were asked to perform was
  - "unclear"/"clear"
  - "easy"/"difficult"
  - "simple"/"complex"
  - "unfamiliar"/"familiar".
- The search that I have just performed was
  - "stressful"/"relaxing"
  - "interesting"/"boring"
  - "tiring"/"restful"
  - "easy"/"difficult".

The average user responses are shown in Table 5, with the most positive response for each user group shown in bold. The figures in Table 5 illustrate that for the bulk of differentials the novice+ users gave the most positive answers. The general tendency, with the exception with how familiar the users were with the task, was that the expert users give the least positive response and the novice+ users give the most positive result. In the case of all of the differentials, with the exception of the clear criteria, the novice+ users give a more positive response than the novice users. It could be interpreted that the expert users are more critical of the tasks and searches, as they have a more in depth knowledge. In contrast to the expert users the novice users are new to the system and are unfamiliar and confused by the tasks and searches; however, when the novice users come back as novice+ users they have already seen the collection and interfaces. Thus all of the user groups have different perceptions

and opinions on the interfaces and search process, meaning that they offer different perspectives.

A multi-way ANOVA was applied to each differential across all systems, all user types and all tasks to test the significance of these differences that have been seen in Table 5; in particular we are interested in the effect of different user types. In the case of familiar ($F = 4.9855$, $p = 0.0089$), restful ($F = 3.8195$, $p = 0.0258$) and easy ($F = 3.7699$, $p = 0.027$) the differences between the users are statistically significant. In a number of other cases it was found that the difference with respect to the tasks was statistically significant; however, we choose to not report these findings here as our focus is on the difference in user groups.

4.3.2 Retrieved videos

In post search task questionnaires subjects opinions were solicited about the videos that were returned by the system. The following Likert 5-point scales and semantic differentials were used. Some of the questions used are contradictions and some of the scales were inverted to reduce bias. The scales and differentials were

- "I had an idea of which kind of videos were relevant for the topic before starting the search" (Initial Idea)
- "I found it easy to formulate queries on this topic" (Start)
- "During the search I have discovered more aspects of the topic than initially anticipated" (Change 1)
- "The video(s) I chose in the end match what I had in mind before starting the search" (Change 2)
- My idea of what videos and terms were relevant changed throughout the task" (Change 3)
- "I am satisfied with my search results" (Satisfaction1)
- "I believe I have seen all possible videos that satisfy my requirement" (Breadth)
- The videos I have received through the searches were:
  - "relevant"/"irrelevant"
  - "appropriate"/"inappropriate"
  - "complete"/"incomplete"
  - "familiar"/"strange"
  - "unpredictable"/"predictable"

Table 6 shows the average responses for each of the scales and differentials above; the labels after each of the Likert scales in the bulleted list above are used to denote the question in the table. The most positive response across for each user type is shown in bold. It can be seen quite clearly that in general (with the exception of two cases Change1 and Change3) that the experts give the least positive responses and the novice+ give the most positive responses, with the replies from the novice+ users being more positive than the response that they gave when they were novice users.

Table 6 User perceptions of retrieved video for each user type, ignoring system (Higher = Better)

Differential	Expert	Novice	Novice+
Initial Idea*	3.781	4.187	**4.406**
Start*	2.812	3	**3.531**
Change1	**3.094**	2.812	3
Change2	3.375	3.469	**3.687**
Change3	2.844	**3.031**	2.781
Satisfaction	3.281	3.344	**3.75**
Breadth	2.625	2.281	**2.781**
Relevant*	3.094	3.406	**4**
Appropriate*	3.125	3.351	**3.937**
Complete*	2.781	3.312	**3.844**
Familiar	3.594	3.281	**3.75**
Predictable	3.094	2.937	**3.344**

Statistically significant differences are marked with $*p < 0.05$

It appears that as with task and search perceptions that the expert users are once again the most discerning users, as they give the least positive responses and thus they are the most critical of the three user groups. Also of great interest is the change in attitude of the novice users, when returning to carry out different search tasks they give more positive results. The novice+ users are more confident about creating search queries (Initial idea, Start), the results that they find (Predictable, Familiar, Complete, Relevant and Appropriate) and the tools that are provided for them to carry out their search (Change2, Breadth). The difference in all of these categories indicates that the novice+ users are more confident in their knowledge of the collection and also their knowledge of the tools. This is perhaps a signal that the change in performance and behaviour cannot be attributed to a simple gain in expertise using the search tools or a simple gain in knowledge of the contents of the collection.

In order to test the significance of some of these findings a multi-way ANOVA was applied to each differential across all systems, user types and the task to test the significance of the differences; in particular we are interested in the effect of different user types. In the case of the semantic differentials, relevant ($F = 5.4395$, $p = 0.006$), appropriate ($F = 3.9164$, $p = 0.0237$) and complete ($F = 9.2373$, $p = 0.0024$) the differences in user perceptions were statistically significant with respect to the user type factor. In terms of the Likert scales the differences for Initial idea ($F = 2.8156$, $p = 0.06553$) and Start ($F = 2.6771$, $p = 0.07617$) were statistically significant for the user type factor. Some of the differences were also significant with respect to the topics; we do not report this as it is beyond the scope of this paper and also this result should not be surprising as the tasks are very diverse and different.

5 Discussion and conclusions

In this paper an examination of the search behaviour and the search performance of users with varying levels of expertise while searching for video has been presented. The evaluation which was conducted put the behaviour of novice and expert users side by side and in particular looked at the influence that a gain in background knowledge has on the performance and behaviour of novice users. There are a number of interesting and important findings that have been made as a consequence of this study. It was discovered that the expert users have the best performance in terms of the number of videos found, the number of relevant videos that were found and MAP for videos that were found. In contrast to this novice users had the worst performance overall, and when the novice users returned as novice+ users their performance improved. However, what is most interesting is that when these novice users returned as novice+ users the difference in task performance is no longer statistically significant in comparison with the expert users as it had been for some performance measures. This is an indication that with some adequate training the performance of novice users' can be comparable to that of expert users. This finding could be very important for future evaluations of multimedia information retrieval systems and also the development of multimedia search interfaces.

Since the expert users have the best performance for our evaluation, the assumption was made that the expert users form the benchmark when it comes to user search behaviour, in comparison with the other user groups in the evaluation. With this assumption in mind an analysis of the search behaviour of the different user types was conducted. Overall, the observation was made that the change in task performance between novice and novice+ participants coincided with a change in search behaviour when the novice users returned as novice+ users. Indeed, the search behaviour of the users in the novice+ group became more similar to that of expert users. Novice+ users had shorter query terms, submitted fewer queries with examples and used fewer examples when submitting those queries than the novice users. This trend held true for both video retrieval systems evaluated. Improved search behaviour and performance when using the SI would indicate that the novice+ users are using their greater knowledge of the collection in comparison with the novice users to improve their performance. The SI encourages users to utilise their familiarity with the video collection, as the SI consists of the bare minimum in terms of query tools to allow users to search the video collection. On the other hand, a change in user behaviour when using ViGOR is also observed between novice and novice+ users. At the same time as novice+ users are using less group searches in comparison with novice and expert users, the behaviour of novice+ users is closer to expert users in other ways. Novice+ users create more groups, use less image examples and begin to utilise the same search tools as expert users. Also across both systems it was observed that the novice+ users' search behaviour becomes more similar to expert users' and less like that of novice users with respect to navigation, novice+ users play videos and navigate more often than novice users and also use the tooltip functionality to gain information about shots more often. Whilst expert users set the standard for the application of these tools and systems, these new findings indicate that novice users can adjust their search behaviour quite quickly in relation to both the content of the video collection and also the search tools that the system provides. Once again this is an extremely significant finding with respect to the design and evaluation of multimedia search tools, as it demonstrates the ability of novice users to acclimatise and discover how to use a new video search system quickly.

While the actions of the novice+ users alter to be closer to the actions of an expert user, it was found that the attitude of the novice+ users towards the tasks and systems was more positive than that of both the novice and expert users. The expert users repeatedly gave the most negative and critical feedback of all aspects of the search, tasks and interface. In contrast when returning to be novice+ users the novice+ users gave the most positive responses. In fact a number of these differences in attitudes were statistically significant based on the user type. Again this finding should be considered in the future design of evaluations of multimedia information retrieval systems and indeed also for future definitions of what are expert and novice users for these systems.

In conclusion, we have presented a new study of the behaviour of different user types while searching for video. The findings of this evaluation have illustrated some potentially new and exciting findings for the evaluation and design of multimedia retrieval systems and will hopefully lead the way in the consideration for expertise in future interactive video retrieval evaluations.

References

1. Jansen BJ, Pooch U (2001) A review of web searching studies and a framework for future research. J Am Soc Inf Sci Technol 52(3):235–243
2. Hembrooke HA, Granka LA, Gay GK, Liddy ED (2005) The effects of expertise and feedback on search term selection and subsequent learning. J Am Soc Inf Sci Technol 56(8):861–871
3. Christel MG, Conescu RM (2006) Mining novice user activity with TRECVID interactive retrieval tasks. Paper presented at the 5th international conference on image and video retrieval
4. Christel MG (2007) Establishing the utility of non-text search for news video retrieval with real world users. Paper presented at the 15th ACM international conference on multimedia
5. Kintsch W (1998) Comprehension: a paradigm for cognition. Cambridge University Press, Cambridge
6. Mat-Hassan M, Levene M (2005) Associating search and navigation behavior through log analysis. J Am Soc Inf Sci Technol 56(9):913–934

7. Campbell M, Haubold A, Liu M, Natssev AP, Smith JR, Tesic J, Xie L, Yan R, Yang J (2007) IBM research TRECVID-2007 video retrieval system. Paper presented at the TREC video retrieval evaluation
8. Smeulders AWM, Worring M, Santini S, Gupta A, Jain R (2000) Content-based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 22(12):1349–1380
9. Naphade M, S JR, Tesic J, Chang SF, Hsu W, Kennedy L, Hauptmann A, Curtis J (2006) Large-scale concept ontology for multimedia. IEEE Multimed 13(3):86–91
10. Hopfgartner F (2007) Understanding video retrieval. VDM Verlag, Germany
11. Guy M, Tonkin E (2006) Folksonomies: Tidying Up Tags? D-Lib Mag 12(1)
12. Halvey M, Keane MT (2007) Analysis of online video search and sharing. Paper presented at the 18th conference on hypertext and hypermedia
13. Hauptmann AG, Christel MG (2004) Successful approaches in the TREC video retrieval evaluations. Paper presented at the 12th ACM international conference on multimedia
14. de Rooik O, Snoek C, Worring M (2008) MediaMill: fast and effective video search using the fork browser. Paper presented at the CIVR 2008
15. Craswell N, Szumer M (2007) Random walks on the click graph. Paper presented at the 30th annual international conference on research and development in information retrieval
16. Hopfgartner F, Vallet D, Halvey M, Jose JM (2008) Search trails using user feedback to improve video search. Paper presented at the 16th ACM conference on multimedia
17. Fass AM, Bier EA, Adar E (2000) PicturePiper: using a reconfigurable pipeline to find images in the web. Paper presented at the ACM symposium on user interface software and technology
18. Urban J, Jose JM (2006) EGO: a personalized multimedia management and retrieval tool. Int J Intell Syst 21(7):725–745
19. Girgensohn A, Shipman F, Wilcox L, Turner T, Cooper M (2009) MediaGLOW: organizing photos in a graph based workspace. In: 13th international conference on IUI
20. Fogarty J, Tan D, Kapoor A, Winder S (2008) CueFlik: interactive concept learning in image search. Paper presented at the SIGCHI conference on human factors in computing systems, Florence, Italy
21. Campbell I (2000) Interactive evaluation of the ostensive model, using a new test-collection of images with multiple relevance assessments. Inf Retr 2(1):89–114
22. Hauptmann A, Lin W, Yan R, Yang J, Chen M (2006) Extreme video retrieval: joint maximization of human and computer performance. Paper presented at the ACM Multimedia
23. Villa R, Gildea N, Jose JM (2008) A faceted interface for multimedia search. Paper presented at the ACM SIGIR
24. Halvey M, Vallet D, Hannah D, Jose JM (2009) ViGOR: a grouping oriented interface for search and retrieval in video libraries. Paper presented at the ACM/IEEE JCDL
25. Schoffman K, Hopfgartner F, Marques O, Boszormenyi L, Jose JM (2010) Video browsing interfaces and applications: a review. SPIE Rev 1(1)
26. Hsieh-Yee L (1993) Effects of search experience and subject knowledge on the search tactics of novice and experienced users. J Am Soc Inf Sci 44:161–174
27. Holscher C, Strube G (2000) Web search behavior of Internet experts and newbies. Int J Comput Telecommun Netw 33(1–6): 337–346
28. Lazonder AW, Biemans HJA, Wopereis IGJH (2000) Differences between novice and experienced users in searching information on the world wide web. J Am Soc Inf Sci 51(6):576–281
29. Bhavnani SK (2001) Important cognitive components of domain-specific search knowledge. Paper presented at the TREC
30. Duggan GB, Payne SJ (2008) Knowledge in the head and on the web: using topic expertise to aid search. Paper presented at the 26th SIGCHI conference on human factors in computing systems
31. Smeaton AF, Over P, Kraaij W (2006) Evaluation campaigns and TRECVID. Paper presented at the 8th ACM workshop on multimedia information retrieval