Evaluating Novice and Expert users on Handheld Video Retrieval Systems

David Scott, Frank Hopfgartner, Jinlin Guo, and Cathal Gurrin

Dublin City University
Glasnevin
Dublin 9, Ireland
{dscott, fhopfgartner, jguo, cgurrin}@computing.dcu.ie

Abstract. Content-based video retrieval systems have been widely associated with desktop environments that are largely complex in nature, targeting expert users and often require complex queries. Due to this complexity, interaction with these systems can be a challenge for regular "novice" users. In recent years, a shift can be observed from this traditional desktop environment to that of handheld devices, which requires a different approach to interacting with the user. In this paper, we evaluate the performance of a handheld content-based video retrieval system on both expert and novice users. We show that with this type of device, a simple and intuitive interface, which incorporates the principles of content-based systems, though hidden from the user, attains the same accuracy for both novice and desktop users when faced with complex information retrieval tasks. We describe an experiment which utilises the Apple iPad as our handheld medium in which both a group of experts and novice users run the interactive experiments from the 2010 TRECVid Known-Item Search task. The results indicate that a carefully defined interface can equalise the performance of both novice and expert users.

Key words: Mobile Device, Keyframe, iPad

1 Introduction

There has been an evident shift in the way we access online content, with the advent of handheld devices and smart phones we have moved away from the rigid structured nature of desktop and laptop environments and embraced the portability and ease-of-use of mobile devices. The level of ubiquitous, always on access that handheld devices provide results in the average user having access to a WWW of information and a small device with (still) limited interaction capabilities to access it. This rush to handheld is further backed by a recent survey carried out by Morgan Stanley[1], estimating that by 2013 handheld devices will have overtaken desktop systems as the most popular portal to the web. It is now apparent that people are likely to access the web from handheld devices in

[1] http://www.morganstanley.com/institutional/techresearch/

a variety of environments, resulting in a search experience that is significantly more cognitively challenging than it was the case a few years ago when we could assume that a user was accessing a video search engine from a desktop computer.

There has been a lot of research efforts in recent years on the development of video search engines using a myriad of available computing devices [6]. A lot of this research has been undertaken through activities in conferences such as TRECVid[2] and VideoCLEF[3]. These video benchmarking conferences encourage knowledge sharing and publication of video search techniques and support the cross-site evaluation of state-of-the-art systems. Participation in conferences such as TRECVid is open worldwide with participants such as Carnegie Mellon University with their Informedia system [15] and University of Amsterdam with their MediaMill system [13] developing novel systems in recent years. While most of this type of research has focused on desktop interaction, TRECVid participants have recently begun to address the new handheld technologies in their video search engine evaluations. For example, DCU's TRECVid submission in 2010 utilised an iPad interface [2] and evaluated the effectiveness of this iPad video search engine on both novel and experienced video search users.

To this end we have focused mainly on content-based video retrieval and the development of new search techniques that can support mobile device access to digital video archives. We want to keep processing on the mobile device to a minimum and not burden the user with excessive requirements for complex interaction with on-screen elements or detailed examination of result sets to identify if the desired information is contained in the particular video document. We strive to develop a simple interface utilising previous knowledge such as storyboarding of video keyframes, utilising of concepts and similarity search [3, 9, 14] to provide expert searchers with the familiarity of traditional content-based systems while introducing novice users to a new and novel way to search, by hiding the complexity of content-based retrieval operation.

In the remainder of this paper we will describe the video retrieval system used by both our novice and expert users. Following this, we will outline our experiments and discuss the results attained by both NIST and by analysis of the post experiment user logs. Finally, we will draw our conclusions and present possible future work.

2 System Overview

Our system was developed as a native iPad application, incorporating a front-end interface and a back-end web service connected to databases and search engine indexes. This provides three methods of searching to facilitate both our expert and novice users: two primary methods (free-text and context search) and one secondary method (similarity search):

[2] http://trecvid.nist.gov/
[3] http://www.multimediaeval.org/

- **Free Text Search:** The first of our primary search methods supports textual querying over three text indexes; meta-data, automatic speech recognition (ASR) text and a phonetically encoding text retrieval system.
- **Concept Search:** The second of our primary search methods, computer vision models trained via Support Vector Machines (SVM), provides a ranked list of the occurrence of any chosen concept e.g. Person, Vehicle, Computer Screen, Face, etc. from within the video archive.
- **Similarity Search:** Our secondary search method implements a relevance feedback technique that, for any given video document, returns a ranked list of the top fifty most visually similar keyframe representations within the collection.

2.1 Interface

From a user interface (UI) perspective, our goal was to develop a system that was easy and intuitive to use for both novice and expert users, while still allowing the user to utilise the available underlying search technologies. This trade-off between the power of functionality and simplicity of use is a well know design issue. By using an iPad device with a touch screen input and by developing a new interface specifically designed for that device we aimed to strike a balance between functionality and ease of use.

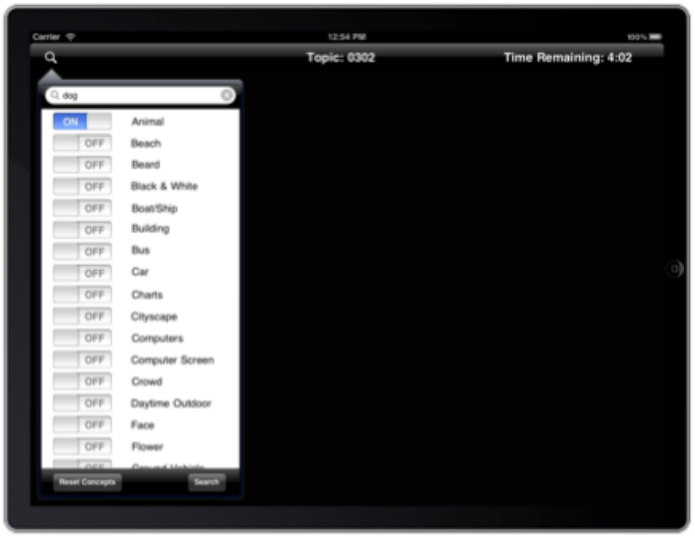

Fig. 1. Search Panel

Upon starting the application the user is required to enter a unique user ID, which allows the system to control the tasks assigned to the user and the

Fig. 2. Search Results

system can then track the progress of the user. Once the user has chosen to start a new topic they are presented with a search panel (as shown in Figure 1). Here, they can input a text query as well as select from a list of 33 predefined semantic concepts. The video results are returned in ranked order to the user: for each video, the title and description as well as a set of keyframes for each shot is shown (the user can scroll to the right to see more for each video). The top ranked shot for each video appears first in the list, with a maximum of 10 keyframes being displayed (selected temporally from throughout the video).

At any point during the search the user can tap on the search icon, which displays the search panel and allows them to refine their search. In addition to this, by double tapping on any keyframe the user can invoke a content-based image similarity search that returns video keyframes that appear visually similar to the one they have selected. After the allocated 5 minutes have elapsed or after the user successfully finds the relevant video the system returns to a topic start page.

2.2 Free Text Search

The text search engine we have chosen to use is Terrier developed by the University of Glasgow [12]. We created three separate indexes over the data and determined a weighting and fusion model by utilising test case topics and results as supplied by the TRECVid organisers.

Source Metadata: Contains metadata information pertaining to the video as crawled from the internet archive and supplied by the TRECVid organisers.

The information stored in this index includes author comments and are generally considered to provide a good overview of each video in the collection.

Automatic Speech Recognition: This index was created by utilising the spoken words in the video and was provided by LIMSI and Vecsys Research [5]. This information was indexed at the shot level by aligning the spoken word to the associated shot bounds.

Phonetic Encoding (PE): PE is concerned with representing the pronunciation of a word with a code made up of letters and numbers [1, 8]. Similar words will have the same code and can therefore be matched by the search engine. Having performed an analysis of several techniques we found that the NYSIIS system [11] was the best choice for this experiment.The output of this process is a set of similar sounding words to the words in the meta-data and ASR which is then indexed by the search engine.

We chose this search engine structure as it increased recall, training topics had revealed that while the average rank fell from 250 to 700 with this index as opposed to a meta only index there was potential to discover 30% more known items with the three index setup.

2.3 Concept Search

Recent research has shown that systems based on the BoW model [7] produced the best results on several large scale content-based image and video retrieval benchmarks.

- SIFT Feature Extraction: The SIFT feature proposed by Lowe has proved to be very successful in applications such as object recognition and image retrieval. To compute SIFT features we use the version described by Lowe [10].
- Construction of Visual Vocabulary: In the construction of the visual vocabulary we employ the Hierarchical K-means algorithm to construct the visual vocabulary based on its advantages of simple and fast implementation. Five million SIFT descriptors were extracted from keyframes from the training data and these were clustered hierarchically using K-means to generate a vocabulary tree with 1296 leaf nodes (i.e. 1296 visual words).
- Visual Vocabulary Transformation: Soft assignment is utilised in the step of visual vocabulary transformation. For each key-point in an image, instead of mapping it only to its nearest visual word, in soft assignment we select the top-100 nearest visual words.

From here we use both positive and negative examples to feed into a Support Vector Machine (see Figure 3) to train the concepts, in the final system we developed 33 concepts based on types of concepts used in the training topics. They are: animal, beach, beard, black and white video, boat/ship, building, bus, car, charts, cityscape, computers, computer screen, crowd, daytime outdoor, face, flower, ground vehicle, in-door, indoor sports, landscape ,map, meeting, military, nighttime, office, outdoor, person, road, sky, snow, stadium, tree, and vegetarian.

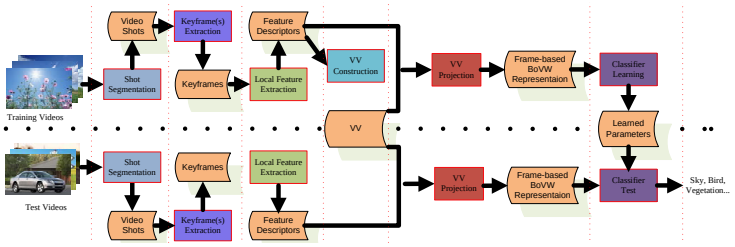

Fig. 3. Concept Training

2.4 Similarity Search

Content-based keyframe search allows users to select a shot on the interface and to find shots visually similar from the collection. For each keyframe in the collection we extracted three low-level MPEG-7 features, namely Colour Layout, Edge Histogram and Scalable Colour.

For each feature we calculated the similarity between each pair of keyframes in the collection. In order to reduce the space requirement for storing the resulting indexes we only stored the top 1,000 similar keyframes for each keyframe. Having calculated the set of similar keyframes for each keyframe in the collection we then combine the scores for each feature into an overall similarity score for a pair of keyframes. For data fusion we we first normalise using MinMax normalisation, using CombSUM[4] to combine the normalised result lists.

3 Experiment

Through our experiments we wanted to compare the performance of novice users against expert users when using a feature-rich content-based retrieval system for video data. In particular we wanted to see if we could develop develop a single search system which could be used by novices and experts alike, with equal performance being attained by both. In addition, we were interested to compare the performance of our iPad search system against other systems taking part in the TRECVid 2010 evaluations. We recruited six users from our research group to complete the task in-house. All of these users had experience working with content-based video search systems and many had participated as users in previous TRECVid experiments completed in DCU, as such this group represents our expert users. We also recruited 12 users to participate from the BI School of Management in Oslo, Norway. None of these users had experience using a sophisticated content-based video search system and none had hands-on experience with using an iPad before. These users represent our novice users. Each participant completed one training topic, followed by 12 search topics during the experiments. We used the Latin-squares experimental design in order to assign users to topics and the ordering of presenting topics to each user was randomised in order to reduce the effects of learning bias, see Table 1.

Novice:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Expert:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**				
Topic 1:	x	x	x				x	x	x			
Topic 2:	x	x	x				x	x	x			
Topic 3:	x	x	x				x	x	x			
Topic 4:	x	x	x				x	x	x			
Topic 5:	x	x	x				x	x	x			
Topic 6:	x	x	x				x	x	x			
Topic 7:	x			x	x		x			x	x	
Topic 8:	x			x	x		x			x	x	
Topic 9:	x			x	x		x			x	x	
Topic 10:	x			x	x		x			x	x	
Topic 11:	x			x	x		x			x	x	
Topic 12:	x			x	x		x			x	x	
Topic 13:		x		x		x		x		x		x
Topic 14:		x		x		x		x		x		x
Topic 15:		x		x		x		x		x		x
Topic 16:		x		x		x		x		x		x
Topic 17:		x		x		x		x		x		x
Topic 18:		x		x		x		x		x		x
Topic 19:			x		x	x			x		x	x
Topic 20:			x		x	x			x		x	x
Topic 21:			x		x	x			x		x	x
Topic 22:			x		x	x			x		x	x
Topic 23:			x		x	x			x		x	x
Topic 24:			x		x	x			x		x	x

Table 1. Table outlining the topic distribution between the novice and expert user groups

The interactive known-item search task at TRECVid 2010 had six teams submit a total of 14 runs. Each run belonged to a certain category depending on the training type and whether the meta-data XML was used or not. For both of our runs we used the meta-data XML (condition: YES) and used only the IACC training data (training type: A). Each system was evaluated based on Mean Elapsed Time, an average of the times recorded for each topic with topics not found being assigned the maximum five minutes. Figure 4 presents the results for all submissions to the interactive known-item search, our two runs are highlighted. Both runs represent results from multiple users where we have picked the best time for each topic in order to populate our submission. Overall our runs came 6th and 7th, however when we compare ourselves against groups with the same condition and training type the position is 5th and 6th.

In the expert run there were a total of 9 topics (out of a total of 22) for which none of our participants found the correct video, interestingly the novice users only missed 8. The fact that users could not find the correct video for certain topics is not surprising, having observed the user experiments it was clear that users found the majority of topics to be either very easy or very difficult. Perhaps more interestingly for us, as part of our post-experiment questionnaire we asked our users to score the system in terms of ease-of-use on a scale of 1–7. For this, our novice users gave the system a median score of 6, with experts giving a median score of 6.5.

Post experiment analysis showed that our novice users executed more searches than our expert users (64 vs 53 on average). From the chart in Figure 5 we see the difference between the users utilisation of the search features; the novice users appeared more reluctant to use both the concept only search (0.88% of the time) and similarity search (5.72% of the time), thus resulting in their requiring on average 9 more queries per experiment (all twelve topics) per experiment to

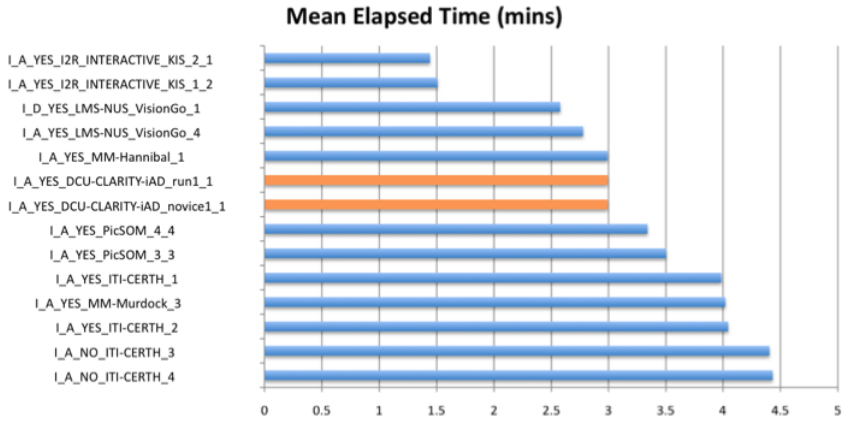

Fig. 4. Official TRECVid Results

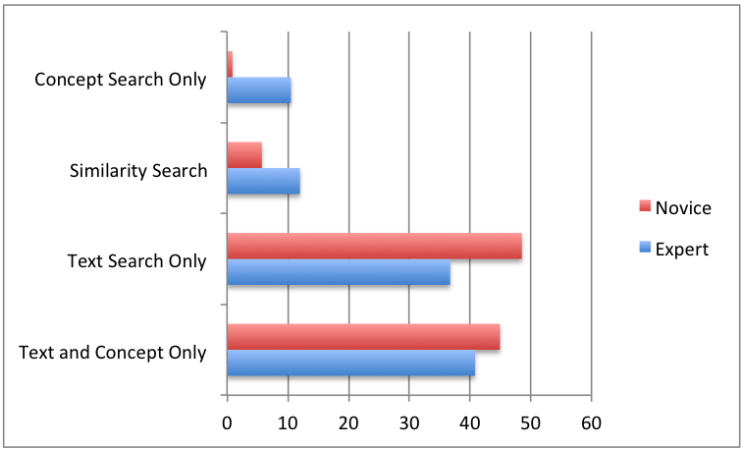

Fig. 5. Log Analysis

attain the same results as our experts. We see from the graph that novice users' preferred search method is text only search nearly 50% of all searches as opposed to experts' who preferred both text and concept search.

4 Conclusions and Future Work

In this work, we developed a video search system aimed at integrating complex search techniques into a single easy-to-use handheld video search engine that can be used equally as effectively by experts and novice users. The results from our official experiments show that the performance of novices versus experts is

identical in terms of mean elapsed time. Through our post-experiment analysis we are investigating why this is the case. One explanation would be that our attempts to build a search engine that could be used by novices and experts alike was successful. Another explanation could lie in the topics used in the search task. Through observations of the experiments we found that both sets of users found the majority of topics to be either very easy or very difficult. The lack of topics of medium difficulty may have constrained our ability to distinguish the differences in performance of different users. Nonetheless through our experimental logs and questionnaires we can still gain valuable insights into the techniques used by both sets of users and their experiences in using our system.

Having analysed the log files further we noted that while the novice users and the experts got the same results according to the official results they relied heavily on text based searches, we intend to further aid these users by incorporating the visual features by using clustering techniques to group like keyframes, this will allow users to more quickly identify relevant keyframes and dismiss a group at a glance thus speeding up their Mean Elapsed Time.

Acknowledgments

The research was funded by Information Access Disruptions, a centre for research-based innovation with CRI number: 174867, funded in part by the Norwegian Research Council.

References

1. Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vassilios S. Verykios. Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19:1–16, 2007.
2. Colum Foley, Jinlin Guo, David Scott, Peter Wilkins, Cathal Gurrin, Alan F. Smeaton, Paul Ferguson, Kealan Mc Cusker, Emma Sesmero Diaz, Kevin McGuinness, and Noel E. O'Connor. TRECVID 2010 Experiments at Dublin City University. In *Proceedings of the 8th TRECVID Workshop*, Gaithersburg, USA, November 2010.
3. Colum Foley, Cathal Gurrin, Gareth Jones, Hyowon Lee, Sinead Mc Givney, Noel O'Connor, Sorin Sav, Alan F. Smeaton, and Peter Wilkins. TRECVid 2005 Experiments at Dublin City University. In *TRECVid 2005 - Text REtrieval Conference TRECVID Workshop*, MD, USA, 2005. National Institute of Standards and Technology.
4. Edward A Fox and Joseph A Shaw. Combination of Multiple Searches. In *Text REtrieval Conference*, pages 243–249, 1994.
5. Jean-Luc Gauvain, Lori Lamel, and Gilles Adda. The LIMSI Broadcast News transcription system. *Speech Commun.*, 37(1-2):89–108, 2002.
6. Frank Hopfgartner. *Understanding Video Retrieval*. VDM Verlag, 2007.
7. Yu-Gang Jiang, Jun Yang, Chong-Wah Ngo, and A. G. Hauptmann. Representations of keypoint-based semantic concept detection: A comprehensive study. *Trans. Multi.*, 12(1):42–53, January 2010.

8. Asjad M Khan, Kathryn S Mckinley, Rotem Bentzur, Daniel Feinberg, Daniel Frampton, Samuel Z Guyer, Martin Hirzel, Antony Hosking, Maria Jump, Han Lee, J Eliot, B Moss, Aashish Phansalkar, Darko Stefanovic, Thomas Vandrunen, Daniel Von Dincklage, Peter Christen, and Peter Christen. A comparison of personal name matching: Techniques and practical issues. In *Workshop on Mining Complex Data (MCD06), held at IEEE ICDM06, Hong Kong*, pages 290–294, 2006.
9. Markus Koskela, Peter Wilkins, Tomasz Adamek, Alan F. Smeaton, and Noel O'Connor. TRECVid 2006 Experiments at Dublin City University. In *TRECVid 2006 - Text REtrieval Conference TRECVid Workshop*, 2006.
10. D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int'l J. Computer Vision*, 60:91–110, 2004.
11. NYSIIS. Comprehensive perl archive network. [online]. *http://search.cpan.org/?krburton/String-Nysiis-1.00/Nysiis.pm*.
12. I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006.
13. Cees G. M. Snoek, Koen E. A. van de Sande, Ork de Rooij, Bouke Huurnink, Efstratios Gavves, Daan Odijk, Maarten de Rijke, Theo Gevers, Marcel Worring, Dennis C. Koelma, and Arnold W. M. Smeulders. The MediaMill TRECVID 2010 semantic video search engine. In *Proceedings of the 8th TRECVID Workshop*, Gaithersburg, USA, November 2010.
14. Peter Wilkins, Tomasz Adamek, Gareth Jones, Noel O'Connor, and Alan F. Smeaton. TRECVid 2007 Experiments at Dublin City University. In *TRECVid 2007 - Text REtrieval Conference TRECVid Workshop*, 2007.
15. Ming yu Chen, Huan Li, and Alexander Hauptmann. Informedia @ TRECVID 2009: Analyzing Video Motions. In *Proceedings of the 7th TRECVID Workshop*, Gaithersburg, USA, November 2009.