PH.D THESIS

Adapting Content Based Video Retrieval Systems to Accommodate the Novice User on Mobile Devices

by

David Scott, B.Sc. (Hons)

School of Computing

Supervisor:

Dr. Cathal Gurrin

September 23, 2013

DCU

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: ________________________ ID No.: ________________________
David Scott

Date: ________________________

Acknowledgments

Firstly, I would like to thank my supervisor Dr. Cathal Gurrin, for giving me the opportunity to work within the research community of DCU and Prof. Alan Smeaton who was always there with an encouraging word. While our group started out small at first, just myself and Sorin Sav, it grew quite quickly. Having a fantastic group of people, with which to bounce ideas off and get feedback/criticism on approaches was great and I would like to thank all the guys from L131 for always being there to iron out issues, chief among these were Frank Hopftgartner, Colum Foley and Peter Wilkins. Another name I cannot forget is Jinlin Guo, working research adjacent we burned the midnight oil on many a project together, a truly wonderful person who I wish the very best of luck when he comes to submit. Lastly from the DCU camp, i would like to thank my friend Mark "Sparky" Hughes, as my tea-buddy he gave me many new perspectives on where and what to research in many of the areas contained within this thesis.

I would also like to thank the Norwegian Research Council for funding my research and the guys up in Tromso who helped me with the direction taken in this research.

My family have been great through this, my mother and father never understood what it is that I do but have always been supportive of my decisions in life and for this I am eternally grateful.

Finally I would like to thank my wife Louise, this last few years have not been easy for her, many nights during the first year of marriage were spent pulling all-nighters to get a project complete on-time, I have a lot of making up to do, but she was always there for me, thank you Louise, my love.

Abstract

With recent uptake in the usage of mobile devices, such as smartphones and tablets, increasing at an exponential rate, these devices have become part of everyday life. This high yield of information access comes at a cost. With still limited input metrics, it is prudent to develop content-based techniques to filter the amount of content that is returned, for example, from search requests to video search engines. In addition, such handheld devices are used by a highly heterogeneous user community, including people with little or no experience. In this work, we focus on the latter, i.e. such casual users ('novices'), and target video search and retrieval. We begin by examining new methods of developing related Content-Based Multimedia Information Retrieval systems for novices on handheld tablet devices. We analyze the shortcomings of traditional desktop systems which favor the expert user formulating complex queries and focus on the simplicity of design and interaction on tablet devices. We create and test three prototype demonstrators over three years of the TRECVid known item search task in order to determine the best features and appropriate usage to attain both high quality, usability, and precision from our novice users.

In the first experiment, we determine that novice users perform similarly to an expert user group, one major premise of this research. In our second experiment, we analyze methods which can be applied automatically to aid novice users, thus enhancing their search performance. Our final experiment deals with different visualization approaches which can further aid the users.

Overall, our results show that each year our systems made an incremental improvement. The 2011 TRECVid system performed best of all submissions in that year, despite the reduced complexity, enabling novice users to perform equally well as experts and experienced searchers.

List of Publications

- D. Scott and C. Gurrin “Searching and Recommending Sports Content on Mobile Devices”, MMM Jan 2010, Chongqing, China.
- D. Scott and C. Gurrin “TRECVid 2010 Notebook Paper”, Gaithersburg, MD, USA.
- D. Scott and C. Gurrin “Summarized-Search: The Fusion of Summarization and Search for Mobile Devices”, iHCI 2010, Dublin Ireland.
- D. Scott and C. Gurrin “Challenges and Solutions for handheld Video Retrieval Devices”, CIICT 2010, Wuhan, China.
- D. Scott and C. Gurrin “TRECVid 2011 Notebook Paper, Gaithersburg”, MD, USA.
- D. Scott and J. Guo and F. Hopfgartner and C. Gurrin “Clipboard: A Visual Search and Browsing Engine for Tablet and PC”, MMM 2012, Klagenfurt, Austria.
- D. Scott and J. Guo and Yang Yang and Kevin Mc Guinness and Zhenxing Zhang and C. Gurrin “Video Browser Showdown Demonstrator”, MMM 2013, Huangshan, China.
- D. Scott and J. Guo and C. Gurrin “Evaluating Novice and Expert users on Handheld Video Retrieval Systems”, MMM 2013, Huangshan, China.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the last decade, there has been an exponential rise in the amount of user-generated content (UGC) populating the web (Cha et al., 2007), ranging from wikis and blog postings to photograph and video uploads. This information explosion is largely attributed to the ease at which one can become a content publisher and has led to an overwhelming array of diverse data. This data is unstructured in nature and differing from professionally edited content, which is usually shot with expensive equipment in a controlled studio environment. One of the most prevalent issues with this new content type is with regard to display, the diverse nature and sheer volume of content make it difficult to determine. Users are presented with information overload in respect to theme and topic of content. It is here we must develop methods which allow the data to be effectively categorized (Smeaton et al., 2006a).

Our research focus as such is on Video Retrieval. In its simplest form this type of retrieval uses external textual evidence, provided at upload, to create a ranked list based on a text ranking algorithm. More advanced systems incorporate methods from machine learning and signal processing, to user feedback and classification, to provide semantic meaning for this content type. The video content itself is dynamic in nature with each video featuring different durations, shots and visual features. Content can range from the short single shot opinion blog to web based video serials. Web resources such as YouTube[1] and Facebook[2] provide a platform for publishing this multimedia content. These sites allow users to publish video content in a social setting, where each video post can be shared among friends and made available in real-time.

Video search/publishing has become part of everyday life for regular internet users, YouTube alone has over 800 million unique visitors each month consuming over 4 billion hours of video content, as of October 2012[3]. This growth in the amount of video data is astounding, processing in 24 hours more video content than has been published in the past 60 years of broadcast television in the United States.

Conferences such as TRECVid[4] emulate the content of these large scale commercial platforms on a smaller scale and are particularly tuned for research. In this thesis we adapt this framework and the task of known-item search to evaluate our techniques in a valid experimental setting. This task models a real world instance where a user is looking for a single item known to be in the collection based on memory of the content. The content provided through this framework features a mixture of both user-generated videos and professionally published content such as news reports and documentary snippets. We have developed a retrieval system over this data which is evaluated against other state-of-the-art systems from within the greater research community.

Our approach to development is to optimize interaction of handheld device interfaces and analyze features which work best in a user testing environment. A handheld device was chosen as the targeted device, as at the stage of TRECVid 2010, these devices were new and yet unexploited technology resource which

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

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

[3] http://www.youtube.com/t/press_ statistics

[4] http://trecvid.nist.gov/

targeted users in a non formal setting. We believed this device best to develop for novice users as they it features simple interaction and required little in terms of learning curve. We began by designing a pilot system to determine appropriate visual features which can aid a set of both novice and expert users. Further experimentation fine tunes and evaluates methods which can increase the likelihood of the user finding specific known-items. We show that by utilizing targeted content based techniques we can enhance the novice users search performance both in terms of reducing the number of searches performed and increasing the number of items found per test.

1.1 Motivation

Content based video retrieval has been a major area of research which has built upon and extended discoveries in the text retrieval world. In the beginning author text associated with a the video is indexed in much the same way as a text document would in a text retrieval system. This however led to many issues which needed to be resolved.

- How can we find video content based on what is visually happening, rather than what is being spoken about or what has been annotated? The spoken word does not always represent the on screen content at a given point.

- Unlike web pages which are just based on text, two sources of data, both video and audio, could answer user queries and both of which should be taken into account when generating ranked lists. Are there ways to understand the content and use this to create more meaningful queries? Can we link videos based on visual features thus clustering similar videos?

- How do we represent the video to an end user with an IR system? Videos are dynamic, not like pictures and text. Can a single keyframe represent an entire video document? Videos are of different lengths, quality, production levels and of differing types. How can we best represent such a varied type of content in a ranked list?

To further advance methods of video retrieval the National Institute of Standards and Technology(NIST) launched a new task in 2001 which focused on video. NIST had long been running a Text REtrieval Conference (TREC).. Two years later, attributed to high levels of participation, a separate evaluation framework TRECVid was founded. Research carried out by participants has led to many advances in shot-boundary detection, keyframe extraction, similarity searching and training of concepts. Some of these features such as keyframe extraction and shot-boundary detection has been implemented in industry standard video retrieval systems such as YouTube.

In 2005, the trio of Steve Chen, Chad Hurley and Jawed Karim founded YouTube. It has become the biggest forum for publishing and sharing video documents in a social environment. In 2011 YouTube had over 100 million unique viewers watching 14.5 billion streams, a 45% increase on previous year according to Nielson[5]. For the most part YouTube consists of private user content, although media companies such as BBC are becoming major publishers due to the scope and audience which YouTube has exposure to. YouTube has also expanded into the mobile market and has been a constant application in all of the modern smart phones such as iPhone and Android since their inception.

In YouTube, the ranked list is determined in much the same way as a search engine where not necessarily the most appropriate video is selected, but one based on the words in textual representation and popularity of the video document. Meta-Data is mined from user comments and annotations at video upload. There is however an inherent problem with this, in that it requires the user to explicitly

[5] http://blog.nielsen.com/nielsenwire/online mobile/number-of-americans-watching-mobile-video-grows-more-than-40-in-last-year/

give information pertaining to the content of the video. Retrieval systems such as YouTube only show a brief representation of the video document in the collection, keyframe and text. From a working background, these video retrieval systems utilize text based retrieval over an index to allow for fast user searching. These retrieval systems also rely on user feedback techniques; so both implicit and explicit capture are implemented. Explicit feedback is logged by users either liking or disliking a video. Implicit feedback is achieved by evaluating what the user is watching with respect to a particular query. This leads to global recommendations such as most viewed and allows YouTube to filter based on user profiling.

In terms of research, benchmarking conferences such as TRECVid try to move away from this traditional retrieval and move more towards what is known as Content Based Multimedia Information Retrieval (CBMIR). This type of retrieval uses both visual features and audio features in conjunction with text and feedback to determine the most appropriate video document for a given query. In the more recent submissions at the TRECVid conference, we see a movement from single keyframe representations to that of storyboarding keyframes as introduced in DCU's TRECVid submission of 2003 (O'Connor et al., 2003). Here keyframes are displayed in a time aligned fashion where each keyframe represents a segment of each found video in the collection.

With the ever increasing availability and uptake of both smart phones and tablets, users are presented with easier on-the-go access to information. A review in 2010 by Morgan Stanley[6] estimated that by 2013 mobile will have overtaken desktop as the most popular framework to access the web. Not only do these devices consume content, they enable users to capture and publish data with ease wherever the user is located. For example, users of the BBC News application can

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

record content on their devices they deem newsworthy and through the native application submit this content directly to the News Desk for consideration.

So how do we make sense of this information? While more experienced users will be able to formulate complex queries to whittle down the results, what of the standard novice user? These users need to be assisted in their efforts to find relevant information, we must develop content based search techniques to better support their information needs.

At present mobile devices have comparatively low processor speed/available power, and limited display size. Visual space and excess results are costly on these platforms. While devices such as the iPad feature a bit more visual real-estate the basic limitations still remain. It is our conjecture that information needs to be as concise as possible to aid the user in locating and interacting with information of relevance. As volume of content rises we will need ways of grouping like content, otherwise users will suffer information overload, scrolling through multiple videos with the same theme or topic. It is due to this content overload and the trending towards mobile access that motivates this thesis research.

In this work, we have developed retrieval systems which bridge the gap between pure research systems which focus on complex querying and the consumer level systems with sparse feature usage outside of text and user profiling. We achieve this by incorporating visual aids seamlessly into queries generated by our casual search user, and displaying the results in a manner which helps users speedily accept/dismiss content. These methods will be explored in detail in future chapters.

1.2 Scope

Within the scope of this research, we define novice users as users who are generally familiar with search systems such as Google and video retrieval systems such as

YouTube. However, they have no formal experience with experimental research systems which incorporate visual features. These systems are commonplace in video benchmarking conferences such as TRECVid. In contrast we define expert users as users who are familiar with both Google and YouTube but who have either been participants or developers in content based search systems in the past.

For this research we focus on handheld devices, specifically that of the Apple iPad. This device has a set display size which is limited when compared with current retrieval systems which can be built over multiple displays and fit more keyframes on their larger canvases. We focus on optimizing the display of content on this specific handheld device, so it can perform as well as similar systems perform on desktop. This work generalizes to comparable devices, but not to the even smaller sized smartphones, which have been left out but are an interesting option to explore for future work

1.3 Objectives, Hypotheses and Research Questions

We have identified two objectives at the core of this research:

1. We aim to demonstrate through the use of real-time video retrieval systems, that designing handheld systems with simple interfaces, using automatic content categorization and grouping, can enable users of different experience levels to attain similar results, thus bridging the performance gap between novices and experts.

2. We aim to evaluate different methods of content representation, both automatically based on visual features and by user testing to define an optimized representation for videos with multiple shots within a ranked list.

1.3.1 Hypotheses

With these objectives in mind we have devised three hypotheses:

1. Using a tailored interface design, which utilizes selected content based retrieval techniques, on handheld devices will increase the performance of novice users when carrying out known-item search tasks.

Our first hypothesis focuses on the design of the content based retrieval system using targeted classification. We assume that by focusing on usable interfaces and specific visual search techniques, we can aid less experienced users in decision making without hampering a more expert search group. To evaluate this we will require the use of a framework which will provide data to create a visual search system.

2. Taking a single keyframe representation approach, where the keyframe is identified by content based techniques, we hypothesize that grouping similar items will help a user to more quickly locate/dismiss relevant videos.

This second hypothesis attempts to address a content visualization approach, by grouping content based on similarity when returned from a ranked list. We believe this will aid user in both easily accepting/dismissing content and quickly finding similar videos without the need for scrolling/browsing. We also believe this cluster list will outperform a method based on the ranked list alone in user testing scenarios.

3. Taking a multiple keyframe representation approach we hypothesize that representing videos in a number of groups will allow for a greater opportunity in finding known-items.

This final hypothesis, attempts to evaluate content when clustered into multiple groups, we hypothesize by dynamically representing content with multiple frames where each video frame belongs to a unique cluster we can increase the likelihood of users finding the known-items. This differs from the previous hypothesis by focusing on the dynamic content of the video, the video can be represented not just once but multiple times, increasing the likelihood of finding known items. The multi-keyframe approach can allow for a single video to represent multiple queries.

1.3.2 Research Questions

In order to evaluate these hypotheses, we must address a number of research questions:

1. Will using a tailored interface design on handheld devices impact performance when compared to other state-of-the-art systems participating in video benchmarking conferences? (hypothesis 1)

2. What visual features will allow our inexperienced users to take advantage of content based search? How will our users interact with the features? How frequently are the features being used? (hypothesis 1)

3. How do we best display to the user an accurate video representation? Is a single keyframe sufficient? Should we use more? Can the automatic use of classifiers help generate this representation? (hypothesis 2)

4. By grouping content can we provide users with a better search experience? Does a ranked list of clusters perform better than a standard ranked list? (hypothesis 2)

5. Should we limit the number of items per cluster group? Should we merge small cluster groups which are visually close? How can we accurately determine this? (hypothesis 3)

6. How can we optimally represent each video and each cluster on the screen of the mobile device while showing clear distinction between cluster groups? (hypothesis 3)

1.3.3 Research Contributions

With the evident move to mobile computing, we have taken steps in adapting technology which was previously rooted firmly in the desktop environment and applied it to handheld devices. In this research we have provided a number of contributions:

- **Implementation of content-based systems targeted at novice users**

Our first contribution in this research is with respect to developing video retrieval systems for less experienced users. Current state-of-the-art systems have varied little since their advent, featuring single keyframe representations and brief text descriptions. This offers the user limited scope of the video document, and relies on user intuition to determine the likelihood of video relevance. The use of content-based techniques can help both with respect to video representation and video search. However, content based systems pose a problem for the less experienced user. They require complex query formulation to attain meaningful results. In this research, we propose a system which is easy to use from the novice user's standpoint but also unobtrusively adapts visual techniques to the user initial query to present them with results.

- **Evaluation of selection criteria for defining keyframe representations**

We focused on evaluating methods of representation, which could identify videos accurately to a user based on a number of different categories. We began evaluating multiple single keyframe representations, using frames both randomly and specifically from the videos, attempting to see if a visual technique could be employed to improve the video representation. Next, we developed methods which focused on the dynamic nature of the video, representing them using multiple keyframes. Each of the methods is user tested, providing user preferences for both single and multiple keyframe representations.

- **Cluster lists which aid users in finding relevant items without the overhead of browsing**

We propose using a clustering approach to represent like content, based on visual similarity. The visually clustered groups allow for content with little or no metadata to be boosted in rank and shown with similar content. From a user standpoint, this reduces the overhead of searching/browsing and allows for acceptance or dismissal of content at a glance.

- **Mobile specific contribution**

While each of the contributions so far do apply to both mobile and desktop environments, it is in the handheld device where the most benefit is seen. These devices have much smaller real-estate space when compared with large screen desktop environments, and these optimization show great potential to enhance the user's search power.

- **Provide a methodology for testing in real world laboratory setting.**

Finally we define a method for evaluation of this system using real users in laboratory settings. We provide an evaluation based on our framework TRECVid (i) Mean Elapsed Time, (ii) Mean Inverted Rank and our own metric (iii) User Search Behavior. Providing a methodology which allows us to retest and evaluate multiple systems with a strict set of metrics, will enable comparative analysis of system improvements made.

1.4 Structure of Thesis

This thesis is structured as six content chapters followed by a concluding chapter summarizing the presented work.

Chapter 1: In this current chapter, we introduce basic concepts of video retrieval, clustering and visual representation, providing motivation for our work, finally presenting our research aims, hypotheses and research questions.

Chapter 2: In this chapter, we provide an in-depth review of research within the area of information retrieval, we pay attention to subsections which directly affect this research, in particular that of video retrieval, keyframe importance and selection, clustering and finally a review of TRECVID.

Chapter 3: This chapter features a technical overview of our prototype system which we use to support our experiments. We present the design, implementation and architecture used in the construction of this system and finally give a high level overview of our intended experiments.

Chapter 4: In this chapter, we determine two sets of experiments. The first features automatic testing with regard to pre-configuration of the prototype system and we determine best text and concepts utilization methodologies. Finally we use this pre-configured system to test novice and expert users and determine similarity between the two user groups. This system will be used to satisfy our first hypothesis.

Chapter 5: In this chapter, we again run two sets of experiments. The first is to determine the best cluster size to distribute the the videos evenly on the interface. After this, we user test a system which utilizes this clustering technique to group content against one which uses no grouping to determine if the user of clustering can indeed aid our end user. In this chapter, we set out to evaluate our second hypothesis.

Chapter 6: Finally, we run our last set of experiments. Firstly we determine the best multiple keyframe representation based on user testing. Next, we utilize what we have learned in previous chapters along with a multiple keyframe representation to allow content to belong to multiple groups. We test this vs a single group representation to determine which will help our users. Our third and final hypothesis prompts the experimentation in this chapter.

Chapter 7: In our final chapter, we critically evaluate the work presented in this thesis with respect to our hypotheses and research questions. We also reflect on the body of work, finally presenting directions for future work.

Appendices: In our appendices we present topics used, experimental guidelines and user survey forms used throughout the user testing phases.

Chapter 2

Related Work

In this chapter, we provide an overview of literature which supports this research. We are concerned firstly with information retrieval, which is a broad and well established research area. Within this field we restrict our view to interactive video search and content representation with a particular focus on handheld devices and we examine relevant information particularly pertaining to our chosen evaluation framework, that of TRECVid.

We begin in the following section with a review of information retrieval specifically that which relates to video. In section 2.2 we give an overview of visual representation methods within the remit of video, specifically that of visual clustering and keyframe representation techniques. Next, we discuss novice users, a major focus of this research, followed by a discussion of data fusion methods which will be incorporated into our system building. Next we discuss TRECVid, a video benchmarking conference with major participation which allows for repeatable experimentation on large scale data-sets and evaluation base on community participation. Finally we finish this chapter with a conclusion by critically analyzing past contributions.

2.1 Background

We begin by giving a historical overview of the research area known as Information Retrieval. Information retrieval is defined as such:

"Information retrieval (IR) is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information."(Salton, 1968)

In general, IR systems are charged with returning items of relevance for a stated information need (Baeza-Yates and Ribeiro-Neto, 1999). This differs from data retrieval, where users attain data from a structured source such as a relational database. IR systems process unstructured data and through translation of the information need can attain items relating to an initial query. IR began in the 1950's, as researchers began to realize that information was being created at a greater volume than was feasible to catalog. As such, there was a need to develop techniques to process this information automatically. Luhn (1957), proposed an approach which used words contained in documents as index terms to create an indexing system. Both the SMART system (Salton, 1971) and Cranfield experiments (Cleverdon, 1997) represent the first experimental IR systems and a methodology with which to evaluate them. Due to this, the 1970's and 1980's featured much work improving retrieval models (Salton, 1971; Salton et al., 1975; Jones, 1980). In 1992 the Text REtrieval Conference (TREC) was founded and funded by the United States government in conjunction with the National Institute of Standards and Technology (NIST) to promote research in the field of IR. Eventually a track (or task) of this conference was set to investigate the viability of IR applied to video. Two years later, due to participation, a second conference TRECVid was formed to further evaluate methods of video retrieval, providing both data and evaluation criteria for developing state-of-the-art systems.

Early video retrieval systems utilized text in much the same way as early text IR systems with description metadata allowing for keyword based retrieval (Smeaton and Quigley, 1996). These early systems evolved and current video retrieval systems are becoming more commonplace as a way to organize our vast repositories of video data, either personal or communal. As such, there has been much research into developing retrieval systems which not only use the features of annotation, meta-data such as title or author comments, but utilize other evidences such as audio and visual analysis of the video document (Blanken et al., 2007). The video benchmarking conference, TRECVid, has provided the data to develop much of the focused research in this area. This thesis will focus on the video retrieval system, particularly that of an interactive system which targets novice users on handheld devices.

2.1.1 Video Retrieval

Video retrieval can be categorized into two separate tracks, as seen in evaluation criteria for TRECVid outlined by Smeaton et al. (2006b). In the first instance, that of fully automatic retrieval, there are no users in the loop. Only one query is posted to the system and a ranked list relating to this query is achieved with items of relevance returned in ranked order. In the second instance, that of interactive retrieval, real world users carry out queries on multiple topics. Unlike the automatic system, users get a representation of the results in real-time and can then, based on these results, reformulate queries to further enhance the rank of the sought after item. See Figure 2.1 for a visualization of both interactive and automatic system designs, for this work we focus on interactive systems. Interactive systems tend to fare better than automatic systems in terms of precision, as automatic systems tend to have items of relevance more spread-out in the ranked list. On the other hand, interactive users can explicitly select the relevant items thus grouping them more effectively (Hauptmann, 2004).

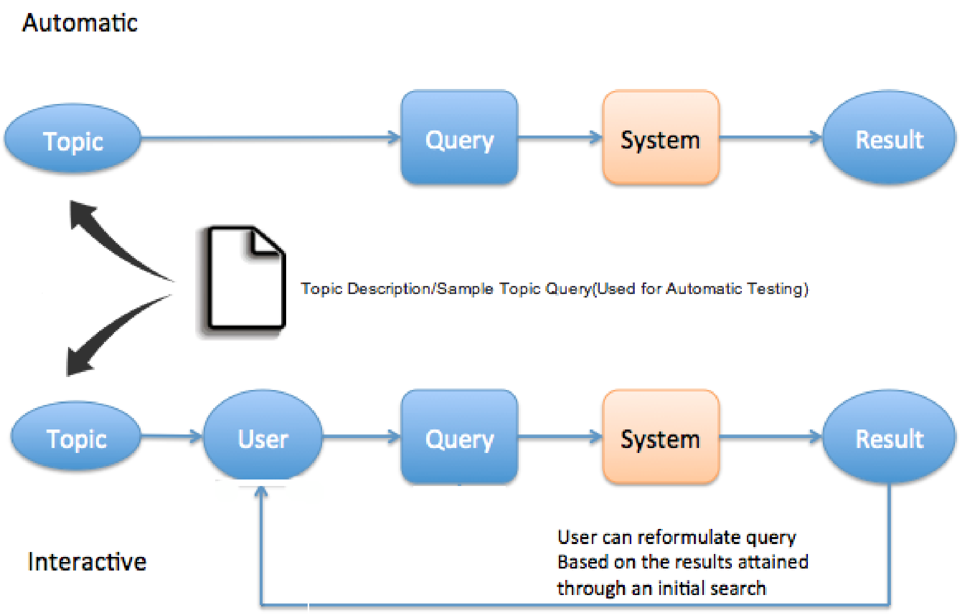

Automatic
Topic
Query
System
Result
Topic Description/Sample Topic Query(Used for Automatic Testing)
Topic
User
Query
System
Result
Interactive
User can reformulate query
Based on the results attained
through an initial search

Figure 2.1: Video retrieval system types

Video retrieval is a complex task, requiring both text retrieval elements and an understanding of the video in order to succeed (Auffret et al., 1999). Video retrieval systems still rely heavily on the usage of text querying to return results of significance. The utilization of visual features to aid in search has had limited effect when compared to that of text retrieval. Most of the mean average precision in standard evaluations can be attributed to text associated with the videos (Hauptmann et al., 2006).

One of the first large-scale experimental systems, the Fischlar video search system, at Dublin City University, encapsulated an end-to-end video retrieval and indexing system over Irish TV and news content, and was an example of a system where content processing leads to more successful result searching (Lee and Smeaton, 2002b; O'Connor et al., 2006). Systems like these, known as Content Based Multimedia Information Retrieval (CBMIR) systems, are complex in nature requiring expert users to formulate queries. This type of system is quite daunting for the average or 'Novice' user. In order to make CBMIR systems more attractive and user friendly to our novice user we must move away from the structured desktop nature of traditional systems and embrace the 'lean-back' (Gurrin et al., 2010) attitude of the modern handheld device. Simplification of content-based features will lead to more successful and rewarding searching strategies. This thesis will study the effects of novice users on a tailored content based video retrieval system through the use of handheld devices. Before we look at mobile device characteristics and the novice user, a short review of digital video processing is presented.

Shot boundary detection is of paramount importance, it allows us to extract the bounds of shots within a video document rather than randomly sampling and either getting too much or too little of the dynamic video information. The shot bounds allow us to extract representative keyframes for each shot, this will aid in our work both in terms of visual representation and with using the representations for clustering.

2.1.2 Structure of Video Data

A video is a sequence of still images run at a specific framerate, such as 25 frames per second (fps), to give the illusion of motion, and can also be accompanied by an audio track. Current video standards which we see in retrieval systems designed through TRECVid, are that of MPEG-1, MPEG-2 and MPEG-4 which are released by the Motion Pictures Expert Group (MPEG), a driving force behind compression techniques used in digital videos. MPEG-1 is used to code VHS standard video. MPEG-2 is an extension of MPEG-1 and as such it provides higher quality audio and video, as seen in DVD. The MPEG-4 standard provides a structure for the storage, transmission and manipulation of the data by representation of video objects which are atomic units of image and video objects (Ebrahimi and Horne, 2000).

MPEG-7 is a ISO standard created by MPEG which provides a unified standard for the description of multimedia data using meta information (Vakali et al., 2004). Various descriptors have been defined to describe visual content, including color descriptor, shape descriptor, motion descriptor, face descriptor and textual descriptor (Salembier and Sikora, 2002).

2.1.3 Shot Boundary Detection

One of the first tasks of TRECVid was that of Shot Boundary Detection (SBD), defined as a process to automatically determine logical boundaries between the shots in a video document. It ran as a track in TRECVid from 2001 to 2007, with more than fifty research groups utilizing a common dataset and scoring metrics in an attempt to solve this problem (Smeaton et al., 2010). Before TRECVid, evaluating methods of SBD was difficult as there were no standardized data sets to be used where different algorithms were subjected to the same content to allow for comparative analysis. SBD was an important content-based task, allowing some structure to be added to the previously unstructured video data (Hanjalic, 2002). This now structured data could be segmented into smaller parts, each part could be analyzed separately to take into account the dynamic structure of the video. Without such a process, interactive search would be a much harder task.

In essence, SBD had two major methods which help in identifying shot boundaries, hard cuts and transitions (Heng and Ngan, 2002; Zhang et al., 1993). Hard cuts are easier to detect. A change happens quite rapidly within the visual structure of the medium, an example of which is by utilizing the edge change ratio (Jacobs A. and O., 2004). Using the low-level feature of edge detection on adjacent keyframes they determined if a major change had occurred and thus a shot boundary had been detected. Transitions are harder to detect as the change happens gradually over a number of frames. Frames either dissolve or fade-in/fade-out (Zheng et al., 2005). For the framework provided through TRECVid participation, we are provided with a list containing the shot boundary of each video, called the shot-boundary master file.

SBD is a key underlying technology that is used in this work to define shots from which we extract keyframes. These keyframes are both presented to the user and form the input for the clustering algorithms. We do not focus on generating new approaches to SBD in this research; rather we take the existing SBD approaches as sufficient for our research and we build on top of them.

2.1.4 Types of Queries

Video data consists of multiple sources such as text, audio and visual features. As such, there are many way with which we can query the video retrieval systems. According to Snoek et al. (2007), three query methods which exist in the domain of video retrieval with which we are concerned with are; text query, visual similarity query and concept query.

Text Query

A popular method of retrieving videos, as seen in sites such as YouTube. As such, novice users have a high level of familiarity with this metric. However, this method relies heavily on text descriptions of the video document. If they do not exist, the video will not be represented by this type of searching.

Visual Similarity Query

Born out of study of image retrieval research, users can query by utilizing exemplar images, from an archive or created through sketching, or a color palate to return items of relevance. it utilizes low-level features such as color histogram and edge detection to determine similarity between content. In the case of video used to find similar keyframe representations, Chen et al. (2000) outline an approach using clusters of images to developing similarity based search systems for both video and image retrieval. This method of search was rarely used by novices.

Concept Query

By building models to create high-level concepts, content can be classified based on a probabilistic score, utilizing machine learning such as Support Vector Machines (SVM). Both positive and negative examples relating to the chosen high-level concept are used to train a successful classifier which can be used as a direct query or as a refinement for other search methods in a video retrieval system. Snoek and Worring (2009) outline a common approach to deploying classifiers in TRECVid style systems. This method of search will be new to a novice user group, having never experienced it in real-world systems.

The Semantic Gap

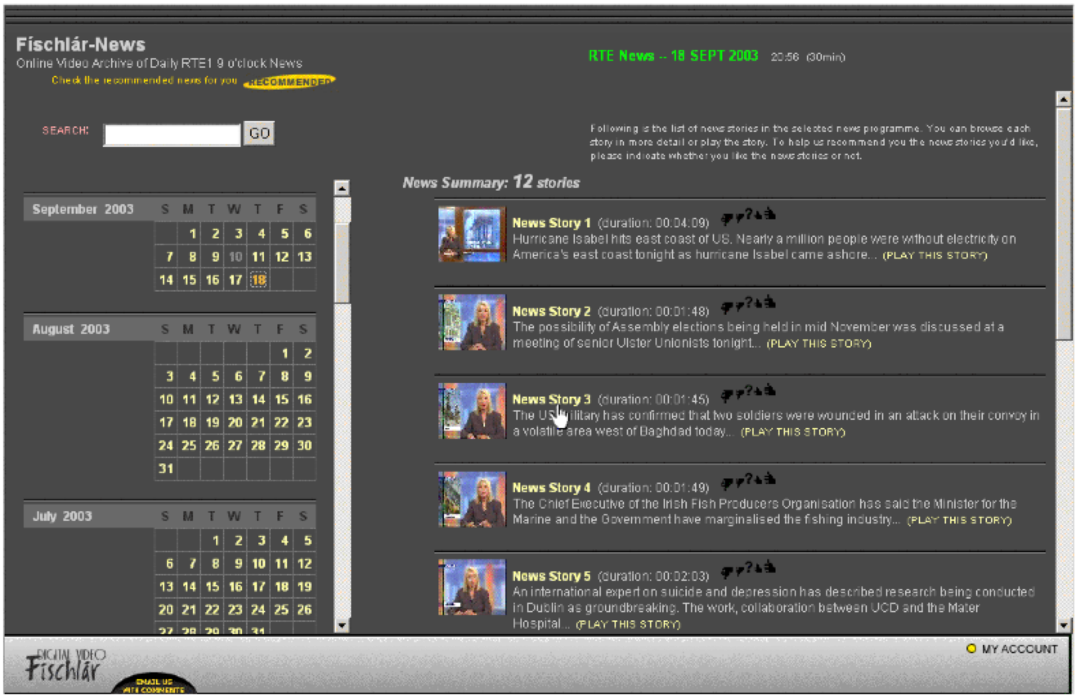
A major problem associated with content-based retrieval is there are discrepancies between the system's understanding of the multimedia data when compared to that of the user perception. This is known as the Semantic Gap and is defined by Smeulders et al. (2000) as:

"The Semantic Gap is the lack of coincidence between the information that one can extract from the sensory data and the interpretation that the same data has for a user in a given situation."

There are difficulties when mapping low-level features to high-level concepts. Some images and video content may, from a low-level perspective, have similar features such as color, shapes and textures but when scrutinized by an end user a clear distinction in the content representations can be seen. One such example of this would be two images from the computer′s perspective containing a blue rectangle on a brown background. According to the low level features, these images are very similar. Upon human examination, it is found that one image is of a blue door in a wall whereas the second image is of a book on a table.

2.1.5 Visual Representation

Visual representation of video data is an important part of the process involved with attaining items of relevance from large scale video retrieval systems. Earlier, we discussed shot boundary detection which allows us to determine representative keyframes for shots within a video. It is how we use these outputs to give a video semantic meaning which is most interesting for us. Lee and Smeaton (2002a) state that the interface performs a vital role in aiding users with validating content. Performance seems to be correlated to the effectiveness of laying out the keyframes to allow users to find relevant items. In Figure 2.2 we see an example interfaces used in the interactive search tasks as part of TRECVid. This however is a very basic design featuring only text search and single keyframe representation of content, though there is some content based techniques with story segmentation.



September 2003	S	M	T	W	T	F	S
		1	2	3	4	5	6
	7	8	9	10	11	12	13
	14	15	16	17	18		

August 2003	S	M	T	W	T	F	S
						1	2
	3	4	5	6	7	8	9
	10	11	12	13	14	15	16
	17	18	19	20	21	22	23
	24	25	26	27	28	29	30
	31						

July 2003	S	M	T	W	T	F	S
			1	2	3	4	5
	6	7	8	9	10	11	12
	13	14	15	16	17	18	19
	20	21	22	23	24	25	26
	27	28	29	30	31		

RTE News -- 18 SEPT 2003 20:56 (30min)

Following is the list of news stories in the selected news programme. You can browse each story in more detail or play the story. To help us recommend you the news stories you'd like, please indicate whether you like the news stories or not.

News Summary: 12 stories

News Story 1 (duration: 00:04:09) ?
Hurricane Isabel hits east coast of US. Nearly a million people were without electricity on America's east coast tonight as hurricane Isabel came ashore... (PLAY THIS STORY)

News Story 2 (duration: 00:01:48) ?
The possibility of Assembly elections being held in mid November was discussed at a meeting of senior Ulster Unionists tonight... (PLAY THIS STORY)

News Story 3 (duration: 00:01:45) ?
The US military has confirmed that two soldiers were wounded in an attack on their convoy in a volatile area west of Baghdad today... (PLAY THIS STORY)

News Story 4 (duration: 00:01:49) ?
The Chief Executive of the Irish Fish Producers Organisation has said the Minister for the Marine and the Government have marginalised the fishing industry... (PLAY THIS STORY)

News Story 5 (duration: 00:02:03) ?
An international expert on suicide and depression has described research being conducted in Dublin as groundbreaking. The work, collaboration between UCD and the Mater Hospital... (PLAY THIS STORY)

Figure 2.2: DCU Fischlar - an example video retrieval interface

Differing from most video retrieval interfaces (Klaus et al., 2010), we chose a platform of the tablet PC, which can be either an iPad, WebOS or any Android tablet, though in our case we selected the iPad. Mobile devices present challenges for user interface designers due to their smaller and more compact screen size and limited input capability (Dunlop and Brewster, 2002; Hürst and Meier, 2008). This thesis sets out to investigate methods of interface layout which can aid interactive search on next generation mobile devices.

Visualization Approaches

In the beginning visualization for video retrieval systems featured largely single keyframe per video representations. This was extended by DCU in developing systems such as Físchlár (Lee and Smeaton, 2002b) which recorded news programmes from the Irish national broadcaster, RTE, and presented the content visually on screen using both single and multiple keyframe methods. In this case, a two-level hierarchical result set was employed in which the first level represented each highly ranked news story/video result with a single keyframe and a second level where a user could choose to explore within a ranked result video and view multiple keyframes (in this case, all keyframes) organized in temporal order. Físchlár implemented a mobile interface that used a single layer, single keyframe approach to video representation (Gurrin et al., 2006) as was limited by the technology at the time.

Other approaches in visualization such as video skims, feature small segments of the video grouped together and displayed via a keyframe to give a summarization of the video document (Christel et al., 1999). This type of visualization was found best to work with highly dynamic content where a single keyframe representation would mean the loss of information (Hürst et al., 2010). It has been shown in many cases that building a relationship based on content such as displaying extra keyframe representations in close proximity to an initial frame, as well as using similarity to link like content has increased the performance of video retrieval systems (Sav et al., 2006). It is with this content based visualization where we focus work carried out in this thesis by adapting clustering of content to group content semantically.

TRECVid Approaches

Three teams have been heavily involved in this interface level representation, namely the Mediamill (Worring et al., 2007) team from University of Amsterdam, the Informedia (Chen et al., 2009) team from Carnegie Mellon University and the CDVP team at Dublin City University (Foley et al., 2010). In this section we will first look at teams who have participated in TRECVid.

The University of Amsterdam's MediaMill prototype, as seen in TRECVid experiments since 2005, utilizes many functional ways to represent both videos and keyframes. For example in the Cross-Browser (Worring et al., 2007) system, they represent videos on the X-axis, while on the Y-Axis keyframes related to each shot in the selected video are displayed. An extended form of the Cross-Browser, the Fork-Browser allows the use of visual features to be used to represent similar keyframes on different axes (Nguyen and Worring, 2008). We see a representation of the fork browser in Figure 2.3. This interface design works well in a desktop environment as the view is constantly being updated based on the user's browser habits. However, this interface design does not maximize the use of the available canvas, especially in the case of the X browser (see Figure 2.3).

Also, inside of the TRECVid framework, CMU have been putting emphasis on the development of content representation algorithms such as the Rapid Serial Visual Presentation (RSVP) (Hauptmann et al., 2006). This technique is used to rapidly present a series of images in a single keyframe, thus eliminating eye movement and giving the user a rapid overview of the content. They found that, in this way, and leading on from other research (Spence, 2002) in serial visualization,

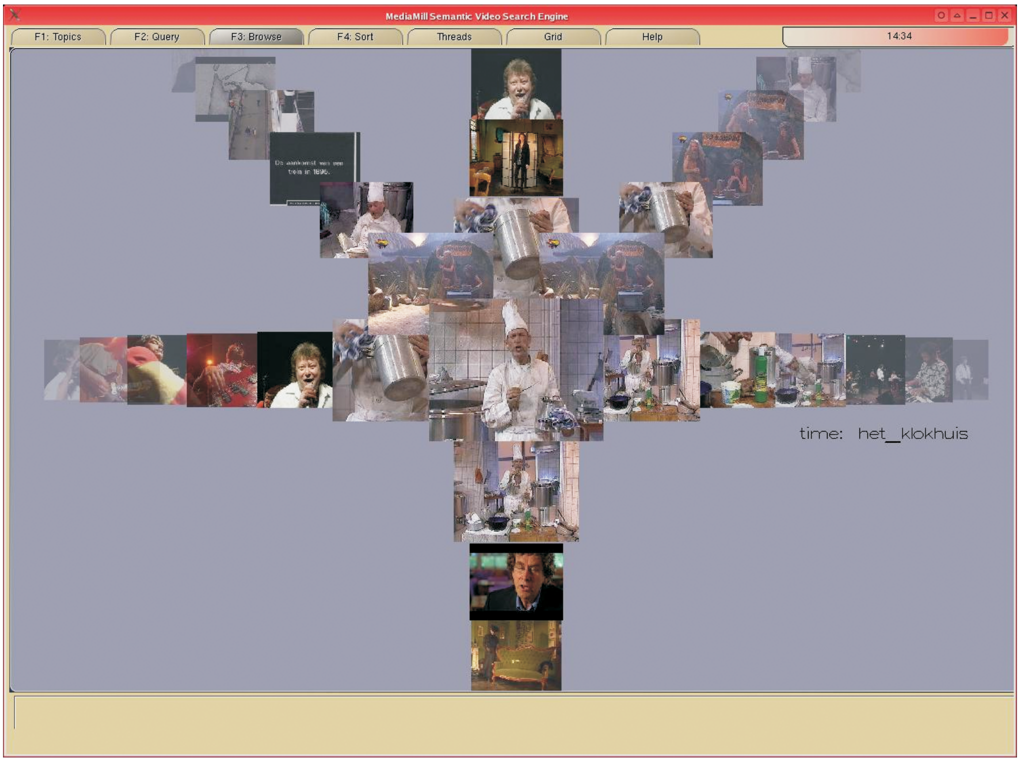
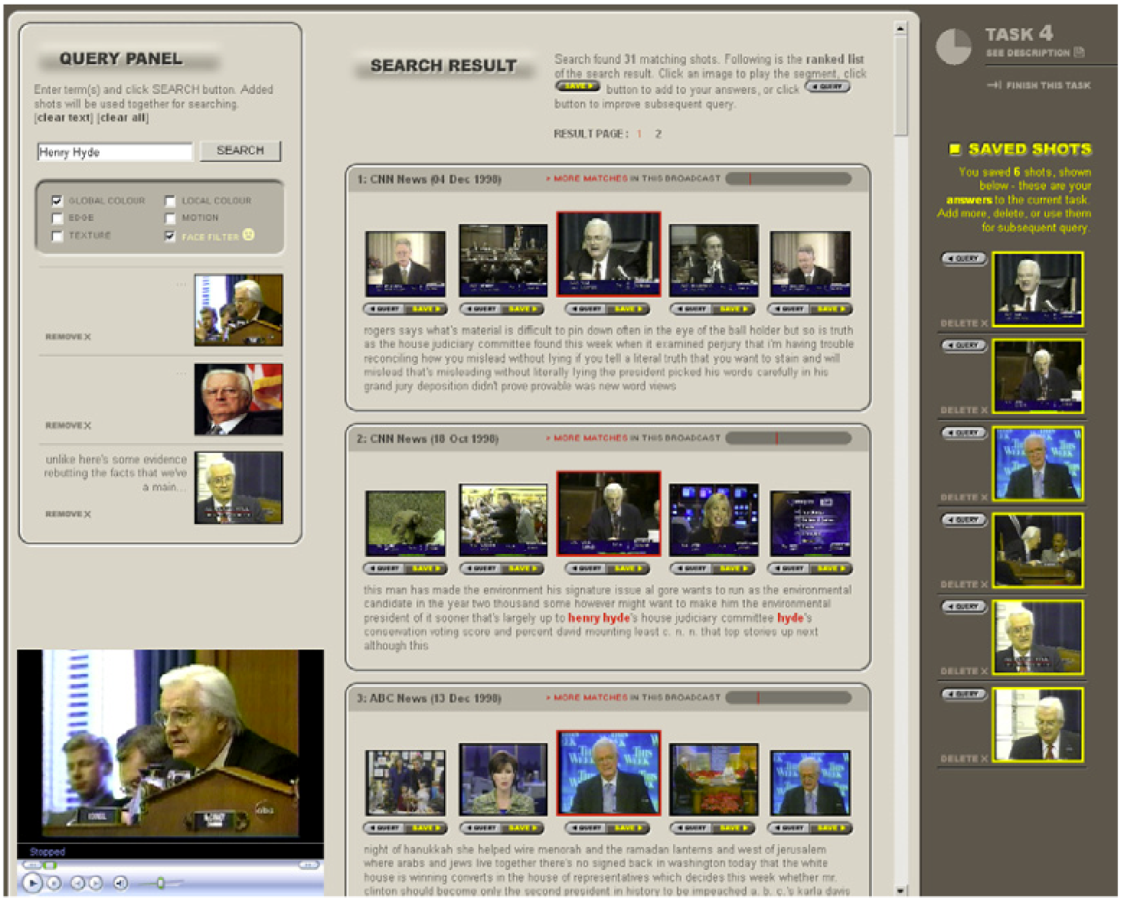


Figure 2.3: MediaMills fork browser, an example of keyframe representation

that users were able to detect a collection of simple images at up to 10 frames per second. Furthermore, they implemented Manual Paging with Variable Page-size (MPVP) which achieved better Mean Average Precision (MAP) as opposed to the standard and stereo RSVP methods. This research focused on only simple images, more complex images require more time to visually understand by the user.

DCU have also employed multiple keyframes representations in its TRECVid experiments in interactive search since 2003 by developing a story-boarding approach to representing each video, see Figure 2.4. With the unit of retrieval being a video shot and not a video, it was determined that if a shot from a given video was positive, shots either side of this shot could potentially be also positive and could at least aid the user in understanding the context of the highly ranked shot. As well as pioneering the storyboard representation, the DCU group has worked on a tabletop interface (Foley et al., 2005), that of the Microsoft Surface, allowing users to co-operatively search to determine shots based on both similarity, concepts and text-querying. This type of interface featured heavily with duplicate keyframe representations due to oversampling of shots, this was part of the task used in identifying multiple instances. For modern retrieval, however, this is a bad strategy. Duplicate keyframes can confuse users and make them think that the system is flawed, using content processing techniques we will eliminate duplicates and give users a better search experience with multiple keyframes.



Enter term(s) and click SEARCH button. Added shots will be used together for searching.

Henry Hyde

unlike here's some evidence rebutting the facts that we've a main...

Search found 31 matching shots. Following is the **ranked list** of the search result. Click an image to play the segment, click SAVE ▸ button to add to your answers, or click ◂ QUERY button to improve subsequent query.

rogers says what's material is difficult to pin down often in the eye of the ball holder but so is truth as the house judiciary committee found this week when it examined perjury that i'm having trouble reconciling how you mislead without lying if you tell a literal truth that you want to stain and will mislead that's misleading without literally lying the president picked his words carefully in his grand jury deposition didn't prove provable was new word views

this man has made the environment his signature issue al gore wants to run as the environmental candidate in the year two thousand some however might want to make him the environmental president of it sooner that's largely up to **henry hyde**'s house judiciary committee **hyde**'s conservation voting score and percent david mounting least c. n. n. that top stories up next although this

night of hanukkah she helped wire menorah and the ramadan lanterns and west of jerusalem where arabs and jews live together there's no signed back in washington today that the white house is winning converts in the house of representatives which decides this week whether mr. clinton should become only the second president in history to be impeached a. b. c.'s karla davis

You saved **6** shots, shown below - these are your **answers** to the current task. Add more, delete, or use them for subsequent query.

Figure 2.4: DCU interface for TRECVid 2004, using multi-keyframe representation

2.1.6 Mobile

Mobile devices present an ideal platform to develop Information Retrieval systems. Given their small size, portable nature, and persistent connections to the world wide web, they provide a gateway to a wealth of information (Tsai et al., 2010). These devices are especially tailored to the inexperienced user. Where complex computer systems can confuse, the simple nature and intuitive layout of mobile devices put the user at ease.

Mobile is a relatively new area of research with one of the first mobile video retrieval systems being developed by colleagues here in DCU (Lee et al., 2001). In this work they create a PDA interface which is a reduced version of the desktop Físchlár system. This interface features limited input metrics and only single a representative keyframe per video, as the device was deemed unsuitable for any other representation approach.

Leading on from this work, Gurrin et al. (2006) showed that these devices were still limited when compared to a fully featured desktop environment where keyframes can be represented in their hundreds. Alternative interaction methodologies were employed by Gurrin that take into account the limitations of the mobile device. In this work, we follow the same concept and develop new retrieval approaches tailored to mobile devices, rather than simply migrate the desktop interface and system to the mobile device.

We believe that these devices are still limited today when compared to desktop, despite recent work carried out by Hürst et al. (2010) that has shown promise that in reducing the size of the keyframe representation on mobile devices, users can make informed decisions as to the validity of the content. They also found in related work, that in the case of their dynamic keyframes(Hürst et al., 2011) which use skims to display more semantic meaning about the content, the user performs better than with static frames.

Interaction on mobile is also different when compared with traditional retrieval systems, where the use of touch based operations replace that attained through keyboard and mouse. Much of the work in this area on mobile is more focused on interactions of browsing through the video than through interaction with a full CBMIR system Hürst et al. (2010).

Xie et al. (2008) were one of the first to utilize a multi-modal search approach on mobile devices. They provided both audio and visual search, achieving good overall results with over 90% accuracy on image categorization. The results in this work, however, are quite static, failing to do content processing on the results puts the onus of searching on the user. In this thesis, we focus on supporting a user in searching for known items in a mobile video search system using a multi-modal approach to searching and content processing techniques to group similar items.

2.1.7 Clustering

Another important area which aids in representation is the use of clustering techniques. Many groups have used both visual and textual features to cluster similar videos. Basic clustering on MPEG-7 visual descriptors such as scalable color, color histogram and edge detection are used to give a vector representation of the keyframes from each video, implemented by using k-means to determine cluster centers and distance from each center to the keyframes. The grouping of like videos allows for users to see similar videos within the same line on an interface level. This type of visualization technique is beneficial to mobile device access as it aids in the reduction of user input.

More advanced forms of clustering techniques like those used by Microsoft Research (Cai et al., 2004) groups in China utilize both text and visual features along with link analysis to determine clustering in web-based search results. Spectral techniques are applied to cluster the search results into different semantic categories. For each category, several images were selected as representative images according to their ImageRanks, which enables the user to quickly understand the main topics of the search results. The combination of textual feature based representation and graph based representation actually reflects the semantic relationships between web images. The reorganization of each cluster based on visual features makes the clusters more comfortable to the users.

Within the remit of video retrieval systems, the team at the University of Defense Technology, Changsha, China (Lei et al., 2004) have been working on applications of visual features for many years. In their paper on clustering work they propose a two-level hierarchical clustering to organize and index the content of videos. At the top level, the text feature space is partitioned into clusters, while at the bottom level, each text cluster is further refined by the use of visual clusters.

Both of these methods offer great results in terms of organizing data semantically, though they are quite computationally expensive and as such cannot be used fully with our limited resources. In this work we explore how to support a novice user (because the majority of users are novices) in searching for known-items from video archives on mobile devices. We will now discuss the characteristics of novice users.

2.1.8 Discussion

Current mobile systems are severely lacking in benchmarking conferences such as TRECVid. Mobile interfaces are somewhat of an afterthought and very rarely do these interfaces take advantage of the full power of mobile, featuring reduced visualization and limited inputs. Even outside of TRECVid most mobile interfaces do not attempt to be anything more than browsers, focusing more on the video content than the system interaction. It is our belief that this next generation of mobile devices offer solutions in developing content rich and fully featured CBMIR system.

2.2 Novice Users

One area which we are concerned with is that of content delivery to novice users. These users are representative of real world casual searchers and as such are unfamiliar with content based search. While there is little research in this area with respect to video retrieval an addition to the annual MMM conference, the Video Browser Showdown, shows that this is a worthy area of research which requires further study.

The VBS encourages the development of technically advanced (non-text) based video search and browsing tools and supports collaborative evaluation in a competitive environment. One of the key targets of the VBS is to evaluate how well novice users interact with complex video retrieval systems, through a dedicated novice user live evaluation during the annual MMM conference. In 2013, our DCU VBS video search tool ranked highest for novice user performance.

In TRECVid each year, systems for the interactive search task features teams who test their systems on novice users. In most cases though these user are novices to TRECVid and not true novices as defined in this work. Christel and Conescu (2006) in TRECVid 2005 used a novice group for their interactive experimentation. Similarly to our experiment, they found that the novice group relied heavily on text as a searching resource. They attempted, through training, to aid these users in performing well on the system, unlike our approach where we adapted automatic aids to enhance the users' search capabilities. In this research we spent a lot of effort in locating and working with novice users to evaluate the developed systems.

2.3 Data Fusion

With multiple ranked lists from different sources we require a method to fuse the information in a way which will not lead to one source overriding the others, unless we specify it to. This area of research is important for retrieval systems where there are multiple sources such as text and visual search components. Work carried out by Vogt (2000) outlined methods of fusion which can aid in combining these different data sources. Shaw et al. (1994) defined six approaches to data combination, two of which were most successful, CombSUM and CombMNZ.

2.3.1 CombSUM

CombSUM is defined as the weighted sum of the document's scores in each of the ranked lists in which it appears (Wilkins, 2007). In the case of a TRECVid system, this weighted fusion can be between multiple separate indexes such as the meta-data and ASR indexes, or between the fused text indexes and a visual search component.

2.3.2 CombMNZ

CombMNZ extends CombSUM. This combination operation rewards documents which appear in multiple ranked lists. Those documents that appear in more lists will be weighted higher, thus increasing the opportunity of finding them in the final fused list.

For our chosen experiments we implemented CombSUM, this has historically been the norm for TREC style experiments and there is little variation from CombMNZ.

2.4 TRECVid

Each year at TRECVid, we see many groups participate in a multitude of tasks, from semantic indexing and copy detection to known-item search and Multimedia Event detection. TRECVid has been running since 2001 and has evaluated different challenges (called tasks) such as shot-boundary-detection in the early years to multimedia event detection in later years. Once TRECVid organizers believe that a challenge has been solved it is retired. Typically 4-5 tasks take place every year. One such task that has remained since 2001 is the interactive search task, where video search systems are interactively evaluated using real-world users in an interactive environment. This task, though the subject of minor focus changes and name changes, is now called the interactive known-item search task; an ongoing challenge that has not been adequately solved.

2.4.1 Known-Item Search Task

The official guidelines of TRECVid 2010 (Over et al., 2010) state that the task of known-item search is defined as:

> "This task models the situation in which someone knows of a video, has seen it before, believes it is contained in a collection, but doesn't know where to look."

This task utilizes data sourced from the Internet Archive Creative Commons (IACC), featuring both video data in MPEG-4 format with textual and shot boundary metadata in MPEG-7 format. Topic descriptions are also provided to aid in the evaluation process of the task. An evaluation of systems submitted by participant to the TRECVid 2010 KIS task is outlined in a paper by Chaisorn et al. (2011).

2.4.2 Evaluation

The evaluation criteria of TREC-style systems have been have been well documented. Descriptions of such can be found in the texts by Rijsbergen (1979), Baeza-Yates and Ribeiro-Neto (1999) and Blanken et al. (2007). Previous experiments were evaluated on the criteria of relevance, precision, recall and fallout. However, since 2010 the interactive search task has changed. No longer are the participating groups tasked with finding multiple instances of relevant items but instead must find only a single item per topic description. Metrics such as precision and recall which are based upon the amount of relevant items become redundant with only a single item of relevance per topic. See below for previous evaluation formula.

$$Precision = \frac{|relevant \cap retrieved|}{|retrieved|} \tag{2.1}$$

Given a set of returned items, precision was the measure of the number of items considered relevant.

$$Recall = \frac{|relevant \cap retrieved|}{relevant} \tag{2.2}$$

Recall determines what amount of the total relevant documents were retrieved.

$$Fallout = \frac{|non - relevant \cap retrieved|}{non - relevant} \tag{2.3}$$

Fallout is the inverse of recall, determining the amount of non-relevant documents which were returned based on the total number of non-relevant documents in the collection.

Known-Item Search Evaluation

To facilitate evaluation of this new task, known item search, new evaluation metrics were devised. Two, in particular, were relevant for the interactive experimentation runs, that of Mean Inverted Rank (MIR) and Mean Elapsed Time (MET), which were provided by the community. We have also determined a method for comparison of interactive experimentation for our own purposes, that of Average Number of Searches Performed.

Mean Inverted Rank

This is similar in a way to previous Mean Average Precision (MAP) evaluations on multi shot retrieval models, in that this evaluation occurs over the entire run rather than on a topic by topic level. A score is attained based on the number of topics in which the system has succeeded in finding the relevant known-item, illustrated by the below equation where *i* represents the number of topics being processed. This is calculated as part of the TRECVid framework.

Mean Elapsed Time

MET is a metric which measures the average time elapsed over the entire interactive run when trying to find a known-item, a maximum time of five minutes applies if the item is not found. The goal is to have the lowest MET possible representing a system which has the potential to quickly find an item of relevance.

$$MET = \frac{\sum_{i=1}^{n} TimeSpentOnTopic(i)}{n} \tag{2.4}$$

Average Number of Searches Performed

This method is used to evaluate in-house versions of developed systems for the TRECVid experiments. Judgments are made based upon the fewest searches performed to find the known-item, illustrated by the below equation where i represents the number of topics being processed.

$$AverageSearch = \frac{\sum_{i=1}^{n} NumberOfSearches(i)}{n} \tag{2.5}$$

2.5 Conclusion

In this chapter we presented a high level overview of video retrieval systems and the components which influence their design and development. In Section 2.1 we look at the evolution of information systems, with a historical, examining aspects of interactions with systems and tasks which have been completed and which aid video retrieval. We also look to Visual representation and mobile. Next in Section 2.2, we examined Novice users, one of the motivating elements for this research. In Section 2.3, we examine Data Fusion, a method for combing the multiple data sources which we have available as part of a CBMIR system. Finally we look to TRECVID, focusing on the task of known-item search and then by looking first to previous evaluation techniques and then to those used for this new iteration of interactive search.

This analysis of related work showed that the current state of the art in video benchmarking conferences such as TRECVid do not use mobile as desktop units provide more power and more real-estate. Those that do offer mobile interfaces have systems which feature reduced input and visualization. It is our belief that not only can current systems support an interaction methodology that desktop systems can, but that due to the ease of use we can develop systems which tailored to novice users can perform as well or better than the current state-of-the-art as seen in TRECVid thus far.

Before conducting our experiments, we must specify a system with satisfies the requirements set out in this chapter for our experiments. In the next chapter we detail our experimental system setup with respect to system architecture, interaction design and users.

Chapter 3

Clipboard: A Content-Based System Designed for Novice Users

3.1 Introduction

To facilitate our experiments we first designed a prototype system[1], Clipboard, which was implemented over the datasets for the Known-Item Search (KIS) task of TRECVid for the three year period from 2010 to 2012. We selected a state-of-the-art handheld device, the iPad, as the deployment platform for our user interface, embracing the defining principle of these devices, ease of use (Ebner et al., 2010). We developed a modular back-end which incorporates content based analysis to empower the search services and to address the topics of this research.

We look at the development of our system from multiple perspectives. We begin by defining the scope of the research and describing the chosen task and evaluation criteria. Next we give an overview of the system architecture. Following this, we outline our interface and how our users interact with the system including the feedback which we record. After this, we describe how the system

[1] "We" is used throughout this thesis to honor the input others provided for example in form of discussions and feedback. Yet, the design and development of the prototype system was entirely my own work, utilizing classifiers which were trained by an external source.

will aid in running experiments and in user evaluation, finishing up by concluding this chapter.

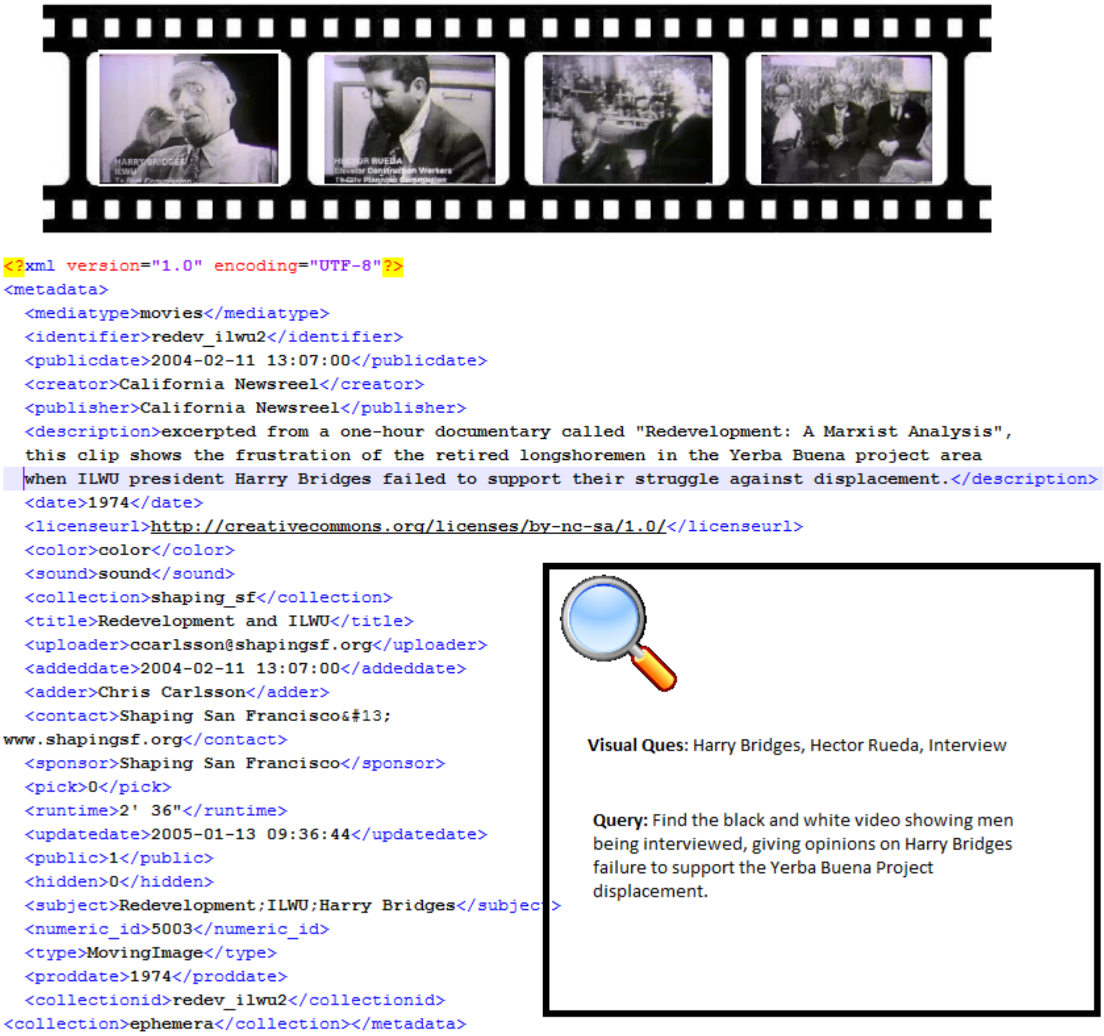
3.2 Known-Item Search Task

The known-item search task began in 2010 as a pilot task of TRECVid in an attempt to model a real world search scenario:

"A user has a requirement to retrieve a video known to be in a collection, the user has no direct link to the video nor the memory of the steps required to formulate the search which initially found the video, the user does, however, have a certain memory of aspects of the content contained within the video" (Over et al., 2012).

The Known-Item Search task was segmented into two types, automatic and interactive search. Automatic systems formulated queries based on the provided test descriptions, utilizing query expansion and visual features to enhance the rank of the known item with only a single search performed. Conversely, the interactive system, by process of human participant involvement, can formulate multiple queries, adapting to search results and drilling down by the use of targeted search terms and visual features. The aim of this task is to locate a single item of relevance as fast as possible. It is in the case of interactive KIS that we focus this research.

TRECVid provided a comparative evaluation framework to develop the Clipboard prototype system which is described in this chapter. This framework contains data in the form of video content, metadata and topic descriptions (Figure 3.1) which can be used to evaluate participating systems. Each year groups were assigned over 200 hours of video data including meta-data traning topics and 24 test topics for interactive and 300 test topics for the automatic test. Standard evaluation such as recall and precision used in previous versions of TRECVid



```
<?xml version="1.0" encoding="UTF-8"?>
<metadata>
  <mediatype>movies</mediatype>
  <identifier>redev_ilwu2</identifier>
  <publicdate>2004-02-11 13:07:00</publicdate>
  <creator>California Newsreel</creator>
  <publisher>California Newsreel</publisher>
  <description>excerpted from a one-hour documentary called "Redevelopment: A Marxist Analysis",
  this clip shows the frustration of the retired longshoremen in the Yerba Buena project area
  when ILWU president Harry Bridges failed to support their struggle against displacement.</description>
  <date>1974</date>
  <licenseurl>http://creativecommons.org/licenses/by-nc-sa/1.0/</licenseurl>
  <color>color</color>
  <sound>sound</sound>
  <collection>shaping_sf</collection>
  <title>Redevelopment and ILWU</title>
  <uploader>ccarlsson@shapingsf.org</uploader>
  <addeddate>2004-02-11 13:07:00</addeddate>
  <adder>Chris Carlsson</adder>
  <contact>Shaping San Francisco&#13;
www.shapingsf.org</contact>
  <sponsor>Shaping San Francisco</sponsor>
  <pick>0</pick>
  <runtime>2' 36"</runtime>
  <updatedate>2005-01-13 09:36:44</updatedate>
  <public>1</public>
  <hidden>0</hidden>
  <subject>Redevelopment;ILWU;Harry Bridges</subjec>
  <numeric_id>5003</numeric_id>
  <type>MovingImage</type>
  <proddate>1974</proddate>
  <collectionid>redev_ilwu2</collectionid>
<collection>ephemera</collection></metadata>
```

Visual Ques: Harry Bridges, Hector Rueda, Interview

Query: Find the black and white video showing men being interviewed, giving opinions on Harry Bridges failure to support the Yerba Buena Project displacement.

Figure 3.1: Example of video keyframes, meta-data and query description for a particular video document used in TRECVid video benchmarking conference

were unused due to the nature of the task and the single item relevance; instead two new evaluation criteria were used:

- **Mean Elapsed Time**: This evaluation metric utilizes the average time taken to complete a topic over the assigned topics in the test. In this way we can determine which system supports the user to locate the correct known item in a faster manner. Lower is better.
- **Mean Inverted Rank**: This evaluation metric is used to evaluate the number of successful known-items found per query in a test session, each successful topic adds one to the running total which is divided by the number of topics assigned to achieve a score between zero and one. A score of zero means no topics were found, where a score of one means all topics were found. Higher is better.

3.2.1 Semantic Indexing Task

The known-item search task parallels with another task, that of Semantic Indexing. This task uses the same data-sets and is responsible for creating concepts which are used to judge whether a visual object is present in given video clips by assignment of a probability score. It is through this task we train our models for Visual Classification (see Section 3.4.4).

3.3 System Description

Our system architecture is composed of separate modular layers, in this respect we can change the behaviour of each layer without impacting on other layers. In this section we describe functional requirements, data used and finally the implementation of the system.

3.3.1 Functional Requirements

To satisfy our needs, we must develop a system with these requirements in mind:

1. *Real-time*: Our system was designed to function with real-time constraints, enhancing the user's search experience by providing instant access to results of search queries. Because it is not ideal to run endless user experiments due to their costly nature in time and finding participants, we developed the system to take advantage of configuration scripts to help us fine tune our modules and weights and measures in our text indexes.

2. *Extensible*: As we mentioned before our system followed a modular design principle; each module performed independently and thus can be replaced without affecting other modules. This means that we can change aspects of the system such as the indexing software or the method with which the visual features are implemented without impacting on the system as a whole.

3. *Feedback and Evaluation*: We designed our system to capture user interaction and task feedback, to provide a view to the search activities of the user. This feedback also captured system interactions and allowed us to simulate real world users in repeat experiments.

4. *Simple and Intuitive Interaction*: Our aim is to make this system as user friendly as possible; this guiding principle will avoid designing complex systems which could potentially confuse our target audience, the ″novice user″. We will explain in more detail when we talk about our interface design in Section 3.5.

	TRECVid 2010	**TRECVid 2011**	**TRECVid 2012**
Hours(approx.)	200	200	200
Videos (Training)	3173	8471	8216
Shots (Training)	119685	144935	137327
Videos(Test)	8471	8216	8263
Shots (Test)	144935	137327	145634
Topics (Training)	122	300	300
Topics (Test)	24	24	24

Table 3.1: Structure of the data for each of the participating years

3.3.2 Data

The data used for the development of our system is sourced through TRECVid[2], arranged under license between the National Institute of Standards and Technology and the Internet Archive. The collection features over 600 hours of both user generated video and professionally edited content. This content represents non-domain specific video and has no defined topic or theme. We are also provided with meta-data pertaining to the video, shot boundary master files and Automatic Speech Recognition(ASR) transcripts. Through the TRECVid bench-marking program this data is released over a three year period. Each year 200 hours of data is released to build and train a system for evaluation.

NIST also provide training and testing topics. These topics are generated by individuals watching video content. Each year 300 test topics are provided for the automatic assessment runs with a subset of 24 being used for interactive experimentation. NIST also provide training topics to aid in the refinement of training systems. This data is outlined in Table 3.1.

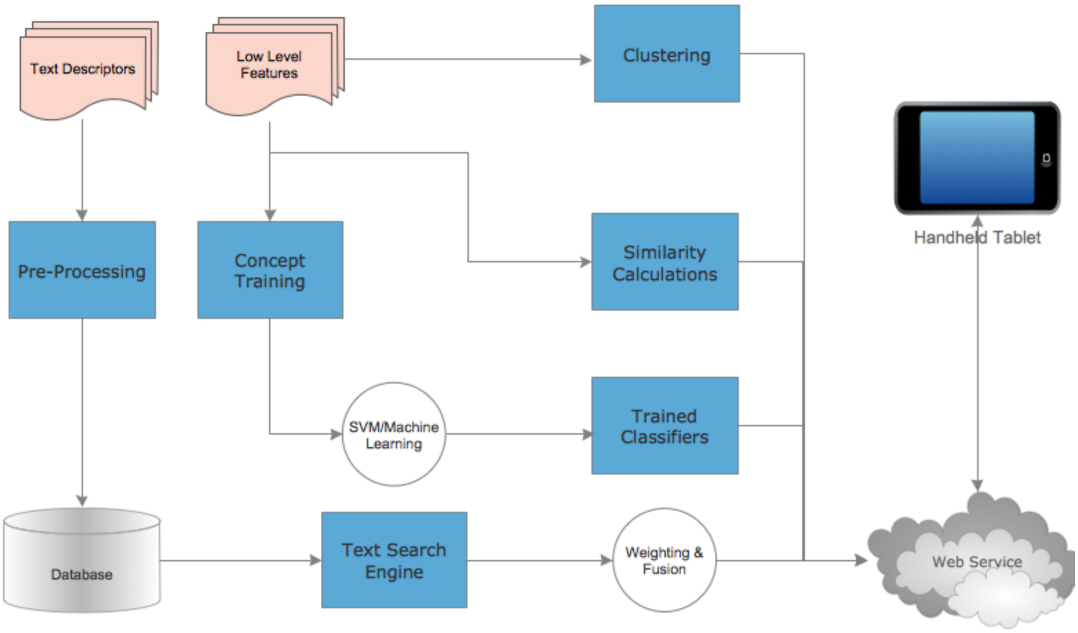


Figure 3.2: Conceptual diagram of Clipboard system

3.3.3 Implementation

The Clipboard system utilizes a data storage layer, communication layer and interface layer which is outlined in Figure 3.2. Clipboard is organized with the following features in mind:

- *Web-service*: acts as the handler from the interface and facilitates our search methodologies.
- *Interface*: a design targeted at handheld device users.
- *Database*: stores the data required to support the text indexes.

[2] http://trecvid.nist.gov/

- *Search Engine*: a self contained unit which encompasses multiple text indexes and utilizes ranking algorithm to return relevant items for each user query.
- *Classifier Support*: trained visual concepts which aid in search and have been used as a primary search also in a boosting and keyframe representation approach.
- *Clustering Support*: groups similar keyframes returned from ranked lists based on the application of a standard clustering algorithm. This module is not used in TRECVid 2010.
- *Similarity Search*: use an exemplar image, attained from primary search, as an input to search. Returns alternative images based on a visual similarity algorithm. This module is only used in TRECVid 2010.

3.3.4 Visual Classification

We will give a brief overview of how our classifiers are trained to better understand how we use them. The first step in the process is to utilize the shot boundary master file, provided though the evaluation framework, to extract the representative keyframes from the training collection. From here we extract visual features which are required in order to train the models. These frames can be represented by sets of the extracted feature descriptors. These sets vary in cardinality and lack meaningful ordering which leads to difficulties for machine learning methods which require vectors of fixed dimensions as input. To combat this we adapt an approach based on the Bag of Visual Words (BoVW) to construct representations for the keyframes.

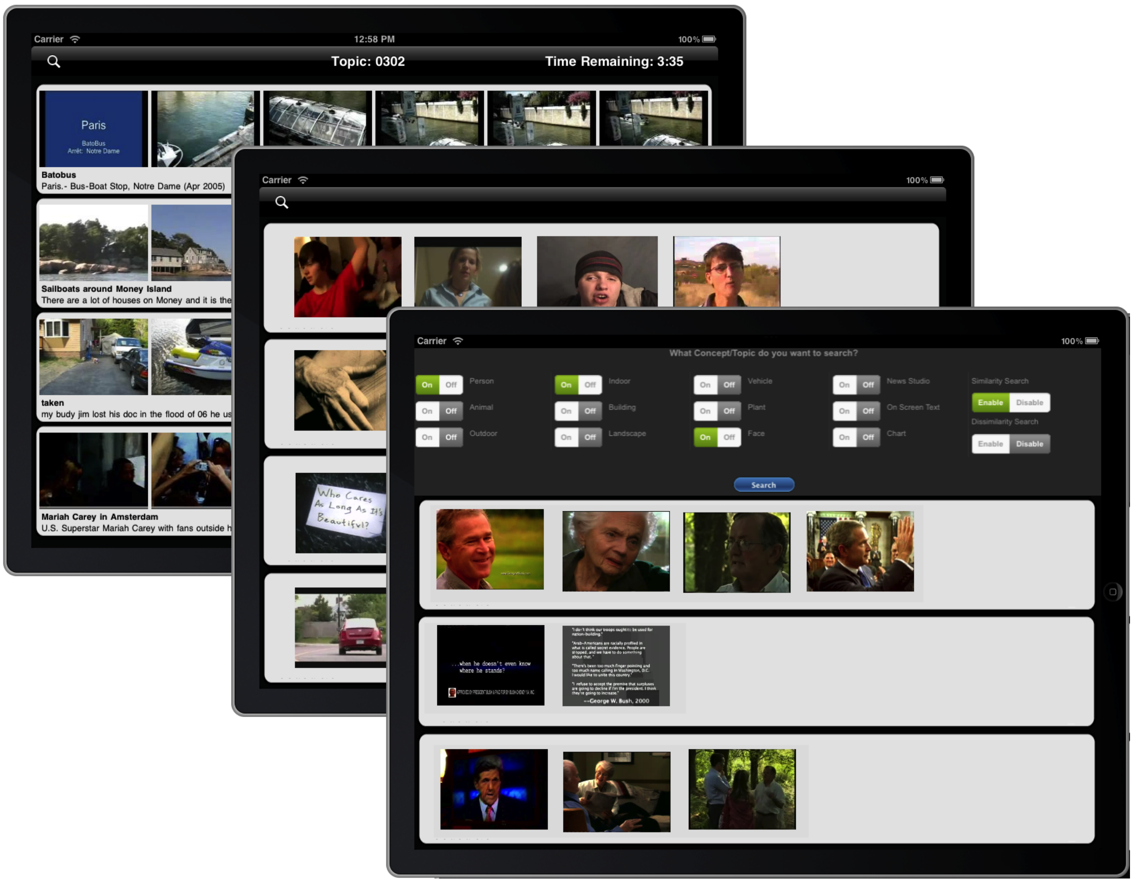
Once images are represented by visual features, we can perform concept detection by using a supervised learning method from labeled images. We adopt a Support Vector Machine (SVM) for concept detection, since it has been proved to be a solid choice, and indeed, it has become the default choice in most concept detection schemes (Tong and Chang, 2001). The RBF kernel, which has been shown to produce good performance (Jiang et al., 2010), is used for the SVM. The SVM classification is implemented using LIBSVM with probabilistic output. Once the models have been trained we apply the keyframes from the test collection, from this we attain a probabilistic score for each of the chosen models based on each keyframe, this forms the basis of our classifiers ranked list.

3.4 Interaction Design

In this section, we will discuss how we developed our system interaction from a user′s perspective. Our aim is to design a system which is both easy to pickup and can aid the user in finding relevant information with a focus on optimal representation of content.

3.4.1 Interface Design

Our interface has evolved from utilizing the iOS SDK in the beginning to embracing HTML 5 quite recently, see Figure 3.3. However, the guiding principle of design has remained the same, simple interaction being key. We have devised two methods of search, the first and more familiar to our target audience is text search. Users can, in much the same way as with YouTube, enter a text based query to attain results which are based on an index match according to certain keywords. The second method of search is by the use of visual features. We utilize the normalized output from the models created during classification training on the test keyframes to create the returned ranked list the users see. Both of these methods of search can be combined through the use of fusion techniques (e.g. CombSUM), allowing us to take advantage of multiple evidences to determine the correct known-item.



Batobus
Paris.- Bus-Boat Stop, Notre Dame (Apr 2005)

Sailboats around Money Island
There are a lot of houses on Money and it is the

taken
my budy jim lost his doc in the flood of 06 he us

Mariah Carey in Amsterdam
U.S. Superstar Mariah Carey with fans outside h

What Concept/Topic do you want to search?

Person: On
Animal: Off
Outdoor: Off
Indoor: On
Building: Off
Landscape: Off
Vehicle: Off
Plant: Off
Face: On
News Studio: Off
On Screen Text: Off
Chart: Off
Similarity Search: Enable
Dissimilarity Search: Disable

Figure 3.3: Examples of the interfaces used by our group for Interactive Known-Item Search 2010 - 2012

Our system lays the results five keyframes wide with vertical scrolling (Smeaton et al., 2003). This method was found to be best when working with dynamic content such as video and we can fit twenty keyframes on each view in landscape mode. This will allow users to easily scroll through results and make a determination of relevance of each item. Prior work for TRECVid also used this layout (Foley et al., 2005).

3.4.2 Feedback

We gather four types of data in this experiment. The survey based feedback documentation can be seen as part of Appendix A:

- **Topic Level Feedback:** User feedback is recorded after each topic, users rate from 1 (Very Poor) - 5 (Very Good) the system performance and 1 (Very Difficult) - 5 (Very Easy) with regard to topic difficulty.
- **Experiment Level Feedback:** At the end of the experiment the users are asked to supply a general overview of how they perceived the overall system. We also get feedback in the form of what the users didn't like, what was confusing and what users would suggest as improvements for our system.
- **System Interaction Logs:** We gather system log data. This will allow us to analyze the user's search behavior, and to evaluate different configuration criteria which would aid the user by re-implementation of their search strategy, if required as a form of automatic evaluation (Foley and Smeaton., 2010).
- **User Profile:** We gather user data, such as age group (18-25, 25+), gender (male,female) and familiarity with both video and information based web search. We also record users' education, though this is not used in our

evaluation due to the group of experts and novices already differing on education level.

To develop our user feedback we began by looking at previous participation in TRECVid, particularly with groups from the interactive search task. We looked at the types of feedback each group captured and compiled a list of criteria which would be appropriate to capture for our system. We added each of these criteria to a form which the users were provided for each test. Each criteria features scalar judgments to be rated on how effective our system was.

The feedback provided users with the ability to inform us of areas in which system improvements could be made, either by directly stating changes or through rating each of the criteria. This feedback drives further research into methods which enhance the search capabilities of the users.

We also capture system interaction logs (see Appendix C), these logs are analyzed post experiment. We evaluate based on number of searches performed, use of classifiers, use of other visual search aids and whether the item was found or not. This information provided a system level view of how each user group interacted with search components and provided insight into methods of improving the system through the log analysis.

3.5 Experimental System Configuration

Our research consists of three user trial based experiments, (i) evaluating search techniques of novice and expert users, (ii) using clustering to enhance the search experience and (iii) evaluation of storyboard clustering and smart keyframe representation. Each experiment requires a different system configuration. In the following sections we will describe how Clipboard supports each experiment.

3.5.1 Experiment I: Expert vs. Novice: A Comparative User Study

We deploy this pilot system to aid in determining features of content based systems which can assist the user in finding items in the known item search task. We begin by utilizing automatic testing to determine different weighting schemes to apply to the text indexes. Next, we evaluate our concepts in both a boosting or filtering technique to ascertain the most appropriate usage. Finally, we integrate these into the final system to evaluate based on user testing. We utilize two metrics to evaluate this experiment, that of Mean Elapsed Time, the average time to complete a search topic, and Mean Inverted Rank, the amount of items found in the experiment over the total items to find. This standardized data will be used to compare to other participating systems and, along with system logs, will aid in determining similarity between our user groups, both novice and expert, described in detail in chapter 4.

3.5.2 Experiment II: Introducing the Cluster-list

For our second set of experiments, we adapt a clustering algorithm to group similar content on this prototype system. We are not concerned with developing state-of-the-art clustering algorithms but instead on how clustering can aid our end users. As such we have chosen the k-means clustering algorithm to evaluate our theory. Our first task is concerned with finding a value of *k* which satisfies our need to have clusters contain approximately five keyframes each, a single row in our interface. This is achieved by automatic testing on training data. The best performing clustering approach is integrated into the test system and user tested against a version which utilizes no clustering to determine which system is best. Our evaluation as with the previous experiment is based on Mean Elapsed Time and Mean Inverted Rank which we will compare with peer groups.

Further analysis of the system feedback will aid in evaluating the best method of representing the content, described in detail in chapter 5.

3.5.3 Experiment III: Visual Representation Comparison

For our final set of experiments we will reuse the clustering techniques attained in Experiment II through this prototype system. This time however, we are testing if a single keyframe representation or multiple keyframe representation is best. Users will run this system with a split of six tasks on the single keyframe clustering system and six on the multi keyframe system. We will evaluate the best approach based on user survey results, as well as the provided Mean Elapsed Time and Mean Inverted Rank, described in detail in Chapter 6.

3.6 Conclusion

In this section we begin by discussing the limitations of the system with regard to the implementation challenges and with the design decisions made. We finish, by giving a summary of this chapter.

3.6.1 Implementation Challenges

As with any system design, numerous challenges arose as part of the initial build. We decided at the beginning of the project to be an early adopter of the next generation of tablet devices. Unlike traditional retrieval systems, these devices have much reduced input with multi-touch as opposed to keyboard and mouse and smaller screen real-estate to deploy content.

One major challenge with the mobile device was with respect to keyframe representation. Initially we displayed all keyframes (up to a certain threshold) of the video, this however led to a huge amount of content on screen, slowing the system in processing the keyframes and making the system less responsive to the users needs. A decision was made to use automatic techniques to alleviate the overhead associated with large numbers of keyframes, outlined in chapter 5 and 6.

Data processing posed another implementation challenge, initially we wanted to process information such as the keyframe data on the device to reduce the overhead associated with loading web based images. Unfortunately, due to restrictions in the initial version iOS for iPad (3.2), there was an app size limit and we could not add all of the keyframe images to the project. This restriction has been since lifted.

Another data processing challenge was with respect to loading the large amount of keyframes from a returned search result. In the first instance we allowed the iOS device to handle the loading of keyframes, this resulted in very slow load times. We used an asynchronous loader to only load keyframes which appeared on the screen to speed up the system. Due to the improvements made through content processing, there were not as many keyframe to load as there was in the initial system and the benefits of this loading were not as apparent as they would have been with larger numbers of keyframes.

3.6.2 Design Decisions

In order to maximize the real-estate space to display keyframe representations an early decision was made to remove the persistent search panel and allow users to explicitly request for search. This allow us to utilize the maximum potential of the device for content display. This is a slight divergence from traditional retrieval and may cause some slowdown with users having to explicitly press a button to reveal a search overlay. We believe that the benefit in screen real-estate far outweighs the discomfort of an extra button press and a non persistent search panel.

Another decision early on, due to the focus on novice users, was with regard to the overall interface design. We did not want the system to be too simplistic thus alienating an expert user group. In the same way we did not want the system to be too feature rich to confuse novice users, leading them to using features but lacking discovery of correct search results within the system. We believe a compromised interface design which was familiar with both novice and expert users were found. Text search at the forefront, with an understated visual search element but overall lacking the ability to formulate extremely complex queries.

3.6.3 Summary

In this chapter, we give a structural and functional overview of the Clipboard system, a content-based information retrieval system designed to be flexible and facilitate our experimentation. We describe the steps involved in both design and implementation, paying attention to interfaces by taking care to keep user interfaces simple and easy for a novice user to understand. We have given a comprehensive description of the system architecture, detailing each component's functionality and how this modular approach allows us more flexibility within the system as a whole.

We also describe how our system supports the experiments, capturing the data required to evaluate the performance. Each experiment allows incremental evaluation of our proposed hypotheses. The above experiments will be evaluated in detail in the following three chapters.

Finally we conclude this chapter by discussing the limitations and approaches to resolution we have faced designing this prototype system.

Chapter 4

Expert vs Novice: A Comparative User Study

4.1 Introduction

In the previous chapter, we presented an overview of the Clipboard system and briefly explained how this system supports our proposed research. In this and subsequent chapters, we elaborate upon these outlined experiments. We deal first with experiments to evaluate the performance of different user groups on simplified handheld Content-Based Multimedia Information Retrieval systems. CBMIR systems have been traditionally developed for expert users with formal desktop environments featuring complex search formulation strategies which are difficult and confusing for casual searchers (Wilkins et al., 2007; Foley et al., 2010). With this in mind, we focus on delivering a system tailored towards this ″novice″ search group. To achieve this, we evaluated a number of potential methods to include in our pilot system. Due to time constraints and difficultly finding participants, user testing was not viable to configure the system. Instead we relied on the use of automatic test scripts, provided through the TRECVid 2010 evaluation framework, to determine an optimized configuration, similar to that performed by Foley and Smeaton. (2010). Our configured system was used to evaluate our first hypotheses:

Using a tailored interface design, which utilizes selected content based retrieval techniques on handheld devices will increase the performance of novice users when carrying out known-item search tasks.

We believe using an interface tailored to both device and user, utilizing visual processing techniques and in a manner which does not bring an overhead to the user, can enhance their overall search experience. To validate this we carried out a real-time user based experiment, with the following research questions used to aid evaluating the hypothesis:

1. Will using a tailored interface design on handheld devices impact performance when compared to other state-of-the-art systems participating in video bench marking conferences?
2. What visual features will allow our inexperienced users to take advantage of content based search? How will our users interact with the features? How frequently are the features used?

For our evaluation we had two user groups, one expert and one novice. Each user in the groups were assigned a set of topics (see Section 4.2), where each topic described a certain video document within the collection. Users were expected to use this topic description to formulate queries to help find the related video, the known-item, within a specific time limit of five minutes. We determined how each group behaved by the use of standard relevance measures, Mean Elapsed Time and Mean Inverted Rank, both provided by the evaluation framework. We also evaluated based on user satisfaction, both on the overall test and on the topic level. Finally we examined the experiment logs to discover search strategies and further compare the user groups.

Within the scope of the experiments we define our two user groups as follows:

- **Expert Users**: An expert user is defined as a user that has a technical expertise within the field of information retrieval, particularly those who have worked with large scale video retrieval systems and are fully aware of the aspects of search. These users understand the fundamentals of visual searching strategies and are able to formulate complex search queries. These users are highly likely to understand search mechanisms.

- **Novice Users**: A novice user is defined as a casual everyday searcher, familiar with text only type searches seen in modern web based search solutions. These users have limited knowledge regarding the usage of visual search techniques. The experiments used in this thesis are their first introductions to systems which incorporate visual search.

In this and subsequent chapters, we detail the data used to implement the system forming the basis of our experiments. We follow this by discussing the two user groups we have recruited to take part in our user testing phase. Next we progress to our automatic experiments used in the configuration of the system. Then we discuss our experiment comparing novice users against their expert counterparts. Finally we close this chapter with a discussion of findings and a conclusion.

4.2 Data

To build and test our pilot Content Based Multimedia Information Retrieval (CBMIR) system we required an appropriate data set. Through participation in the TRECVid Interactive Known-Item Search (KIS) task in 2010, we were provided

Structure of the Data TRECVid 2010 Interactive Known Item Search

Hours Of Video	200 (approx.)
Number of Videos in the Training Collection	3173
Number of Shots in the Training Collection	119685
Number of Videos in the Test Collection	8471
Number of Shots in the Test Collection	144935
Number of Topics in the Training Collection	122
Number of Topics in the Test Collection	24

Table 4.1: Overview of the keyframes/shots used for TRECVid 2010

with an evaluation framework. This framework facilitated our data need by providing data mined from the Internet Archive Creative Commons (IACC). The pilot task utilized the *iacc.1.a* data corpus which features generic videos with durations of up to 3.5 minutes in length. Videos ranging from single shot UGC to professionally edited content were contained in the collection. The framework provides us with over 200 hours of video content. This is detailed in Table 4.1, which is segmented further into two collections, one for training and the other for testing purposes. The contents of the framework are:

- **Meta-data**: gives us information regarding *video title*, *a short video summary*, *keywords* and a host of other details such as *date uploaded* and *video duration*, forming the basis of the data used in our text search index.

- **Text Transcripts**: Through the framework, we were supplied with data in the form of Automatic Speech Recognition (ASR). Each of the videos goes through a machine learning process to extract the spoken word (Gauvain et al., 2002). In the case of foreign language a further machine translation to English is performed. We utilize this data in conjunction with the shot boundary information to align the spoken word to each shot. This data is incorporated into a seperate text index. The return type from this index is shot level, which was aggregated to a video level, the return type of this TRECVid test.

- **Shot Boundary Master File**: This file contains output with regard to detected shot boundaries for each video in the collection (Qunot et al., 2003). Each shot is represented by a start and end time along with a time-stamp for the representative keyframe. This shot boundary file aids in the extraction of keyframes used in both visualisation on the interface level and training content based search techniques.

- **Visual Ground-truth**: Each year at TRECVid to aid visual classification, members of the TRECVid community participate in a collaborative annotation task (Ayache and Quenot, 2008). Each shot in the collection is judged with respect to assigned classification models, examples being People, Buildings, Vehicles and Vegetation. These judgements form the ground-truth for building the models to be used on the test data sets.

- **Topics**: Topics are defined by NIST assessors who watch a sample set of video documents and, without knowledge of the underlying meta-data describe the events of the video. Each topic features multiple visual cues which can help with identifying effective classifiers and a text description of high level actions within the document. With respect to the iacc.1.a there were a total of 422 topics, 122 training topics and 300 test topics of which 24 test topics were used for the interactive search task. Examples can be found in Appendix B.

- **Evaluation Criteria**: Both Mean Inverted Rank and Mean Elapsed Time (see Chapter 3) were used to evaluate the performance of the retrieval system. These methods provide for evaluation of repeatable experiments and for accurate comparisons with participants of the TRECVid video benchmarking conference.

4.2.1 Additional Data Sources

Three additional sources of data with which we use to develop our system are:

- Phonetic Encoded Strings
- Extracted Visual Features
- Similarity

To enhance our probability of finding known items and to combat both spelling errors in the annotations and misinterpretation of aural information by the topic annotators, we utilized a phonetic encoding strategy. Phonetic encoding is concerned with representing the pronunciation of a word with a code made up of phonetic sequences. Similar words will have the same sequence and can therefore be matched by the search engine, forming the data required for our text search index. To avoid numerous false positives associated with phonetic encoding due to the vast number of similar sounding words, automatic testing was used to weight the index to avoid the phonetic index dominating the ranked list (see Section 4.4.1).

We extract visual features from the keyframe representations, JPEG images from the video's middle frame using the shot-boundary master file, both on the training and test collection. In the case of the training set, we feed these visual features into an SVM based on positive and negative concept examples returned from the visual ground-truth. The output of this task is a model based on the positive element of classification. We repeat this process for each of the classification models we require, each keyframe from the test collection is applied to these models to determine a probabilistic score of containing features which satisfy the classifier. In this way we build lists of classifiers, ordered by probability scores, which we use as a data source for the system.

Our final source of data also utilizes the extracted visual features based on the shot-boundary master file. This time however, only the test set is used. We process each keyframe in the collection against all other keyframes using a similarity algorithm based on MPEG-7 features. The output of this algorithm is a list of the top 100 similar keyframe images for every keyframe. We utilize these lists to quickly generate similarity results on keyframe queries.

4.3 Users

For our experimentation we have two equally sized user groups, novices and experts. Each user was asked to fill out a feedback form before, during and after the experiment. From these forms we captured the following demographic information:

- **Expert Group:** Consists of members who have either been directly involved in previous developments of TRECVid systems or have experience in using content based systems. These users have a keen knowledge of how these systems work and have experience with using visual elements of search to formulate complex queries. As such, we have recruited 8 participants from our research group many of which have been directly involved in TRECVid over the years but none who directly influenced the development of this system. The majority of participants hold an advanced degree with only a single research assistant having a bachelors degree. The average age the participant was 27 with a standard deviation of 4.2. Most participants engaged in moderate to heavy video and web searching activities, with this group being predominantly male.

- **Novice Group:** Consisting of participants from a non-technical background, all recruits were undergraduate/masters students attending a business

Participant Profile		Novice	Expert
Age:	18 - 25	4	1
	25 - 30	3	6
	30 - 35	1	1
Web Search	Regular	6	7
	Infrequent	2	1
Video Search	Regular	6	5
	Infrequent	2	3
Education	Undergraduate/No Degree	12	1
	Graduate	0	4
	Researcher	0	2
	Faculty/Staff	0	1
Gender	male	3	7
	female	5	1

Table 4.2: Novice and Expert Participants Profile

course in The Norwegian School of Business Management, Oslo, Norway. Most of these had never used a tablet before. The user video retrieval experience was limited to searches carried out on YouTube and they had no formal know-how in either development or usage of content based systems. We attained 8 volunteer students which we used to run our experiments, with an average age of 25 years with standard deviation of 4.77. These participants engaged in frequent web and video searches, with the majority of participants being female.

We visualize the participants profile in Table 4.2, here we show a side-by-side breakdown of the user groups. By using these two user groups we will be able to evaluate our first hypothesis.

4.4 System Configuration Phase

Before testing with real-world users, we began by carrying-out configuration experiments. This allowed us to, through automated techniques, create our final tailored system. Our first experiment revolved around determining the best weighting on our three text indexes to attain the highest score on the ranked list. Next, we examine methods of applying concepts to the content, first by applying classifiers as filters and secondly by applying boosts to content based on classifier confidence. Finally we examine methods which allow us to fuse both the text and classifiers to form the final ranked list.

4.4.1 Automatically Weighting the Different Indexes

The three indexes created from textual data pose a problem with regard to fusion of the ranked lists. It was necessary to determine a weighting scheme which will provide the best possible rank for our known items. To achieve this weighting scheme we utilized the 122 topics which have been provided as part of the training data set, creating automatic test scripts from these topics. The test scripts were run on each of the indexes providing three separate ranked lists which we normalize using MinMax normalization (see Formula 4.1). We applied different weights to each of these indexes, fusing them using a method of CombSUM (Wilkins, 2007), which fuses by addition of the weighted ranked list values, a strategy which proved best for Wilkins. We ran 125 iterations of weighting with each index receiving a maximum rank of 5. We found that the weighting scheme of 5:2:1 best for the Meta, ASR and Phonetic index respectively. Table 4.3 represents the average rank of the known items in this ideal weight.

$$Norm_{score(x)} = \frac{Score_x - Score_{min}}{Score_{max} - Score_{min}} \tag{4.1}$$

4.4.2 Automatically Optimizing Methods of Classification

Previous CBMIR systems as seen in conferences such as TRECVid have been successful in utilizing classifiers as filters, omitting results where probabilistic

Meta	ASR	Phonetic	Average Rank
4	5	4	849.95
4	5	5	848.75
5	1	1	844.72
5	2	1	***840.65***
5	3	1	841.71
5	4	1	844.45

Table 4.3: Table outlining the average rank of the known item for specific weighted fusion (lower rank best) - This is a subset of the experiments. Integer weighting best supported by Terrier.

Topics	**Filter 0.5**	**Filter 0.75**	**Boost 0.5**	**Boost 0.75**
Topic 1: Find the video of men and women running with musicians on a stage, one with long hair and guitar. **Classifiers Used:** Person, Outdoor	101	101	20	12
Topic 2: Find the video of the show ”Good 2 Know” with Pr.Todd, Youth Pastor, behind a huge fake mouth **Classifiers Used:** Person, Indoor	101	101	101	101
Topic 3: Find the video showing a man wearing glasses speaking French in an interview. **Classifiers Used:** Person, Face	16	7	30	19
Topic 4: Find the video of Kerry and Bush political ads **Classifiers Used:** Person	25	101	30	22
Topic 5: Find the video with a procession of people walking down the street **Classifiers Used:** Person , Crowd	101	101	101	101
Topic 6: Find the video of a woman in a light green dress and huge black hat. **Classifiers Used:** Person	90	54	96	75
Topic 7: Find the video of Humvee truck explosion described in Arabic **Classifiers Used:** Vehicle, Outdoor	30	101	36	31
Topic 8: Find the video of a blonde in pink seated with a coffee cup at hand giving financial advice. **Classifiers Used:** Person, Indoor, Office	101	101	101	101
Number Found	4	2	5	5
Average Rank	70.63	83.38	64.38	57.75

Table 4.4: Rank of known item within the top 100 based on multiple boosting and filtering approaches

scores are below a certain threshold. While this method works well for systems where results are based on multiple shots, we believe that this can prove quite restrictive when looking for a single videos as with the known-item search task. If the single item is filtered out by the classifiers we may find similar items but never find the correct video. We attempted to use concepts in a boosting approach to combat this problem. In the following experiment, we evaluated the performance of our classifiers with both a boosting and filtering approach. We took eight exemplar topics from the training set and created automatic queries featuring text and selected classifiers which we believed would benefit the query. In Table 4.4, we compare our boosting approaches with filtering approaches of thresholds greater than 0.5 and 0.75. We find that the upper quarter boosting technique yields an average rank of 57.75, much better than both filtering techniques. In fact, both boosting techniques far outperform the filtering technique. The advantages of this boosting technique showed that, while sometimes the boosting approach did increase the position in the ranked list, it more often found the known item having found 5 items in both boosting tests compared with 4 and 2 respectively with the 0.5 and 0.75 filtering approach. For the known-item task, we believed that the 0.75 boosting approach would give the best chance at achieving better results in the ranked lists.

4.4.3 Automatic Optimization of Final Fusion

Unlike filtering which functions like an on/off switch, boosting required fusion with the ranked list attained from the text search engine to be effective. As such with two rankings we required a dedicated weighting scheme. Our final set of automatic experiments revolved around determining the best fusion of both our chosen weighted text search and boosted classifier ranked lists, we utilized the topics from the previous experiment where the known-item was found. We used

Text Weight :	**2**	**1**	**3**	**3**	**1**	**2**	**1,2,3**
Classifier Weight :	**1**	**2**	**1**	**2**	**3**	**3**	**1,2,3**
Topic 1:	11	15	3	7	20	15	12
Topic 3:	21	17	24	20	16	18	19
Topic 4:	21	22	20	22	26	22	22
Topic 6:	60	79	52	64	88	80	75
Topic 7:	28	33	26	28	35	32	31
Average:	28.2	33.2	**25**	28.2	37	33.4	31.8

Table 4.5: Final fusion of both text and classifiers

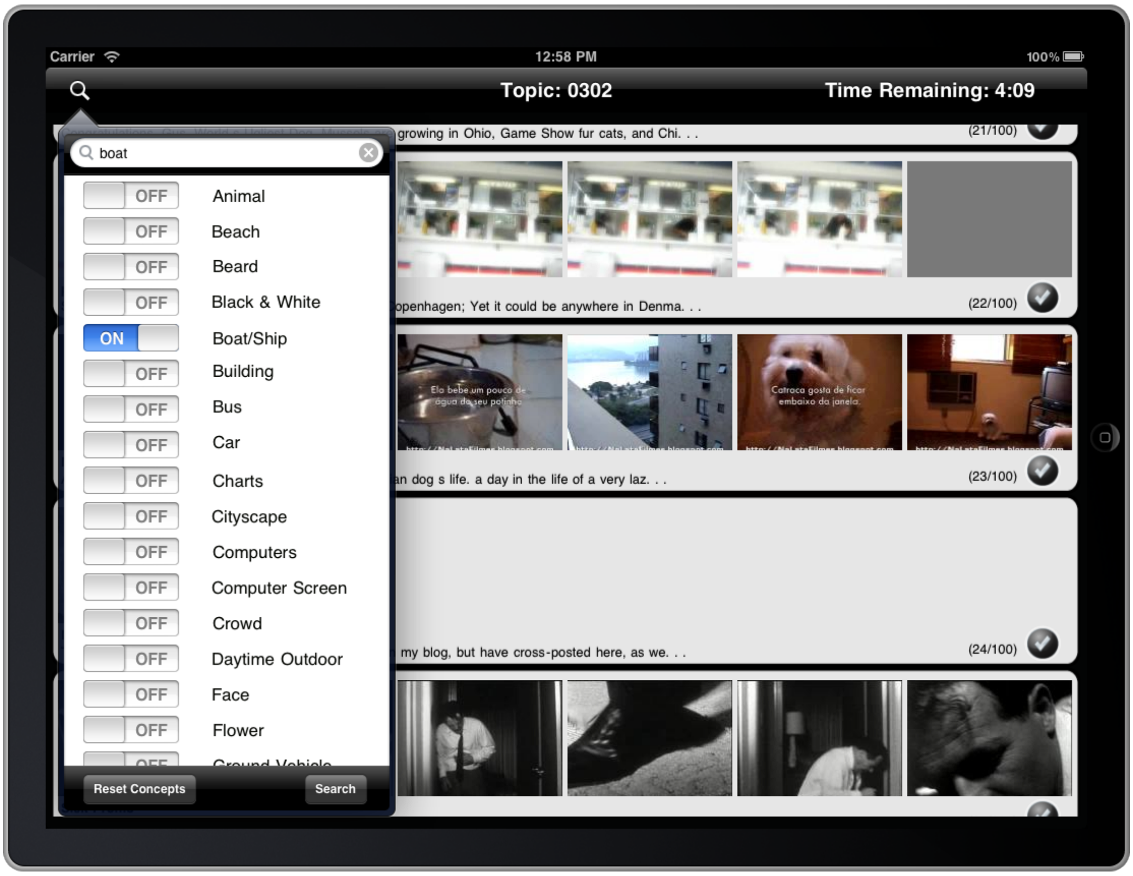
Concepts
animal, person, indoor, car, vegetation, landscape, building, bus, cityscape, boat/ship, computer screen, crowd, face, ground vehicle, military, outdoor, tree, meeting, nighttime, road, sky, office, beard, computers, flower, Black and White video, daytime outdoor, indoor sports, map, charts, beach, stadium, snow

Table 4.6: List of classifiers used in our experimentation

nine iterations to weight both of the normalized lists and found that our best possible weights for both text and classifier to be in the relationship 3:1, see Table 4.5.

4.5 Interactive User Evaluation

For this experiment, we utilized the prototype system outlined in Chapter 3 (see Figure 4.1 and 4.2). We implemented it with the three weighted text indexes and the 33 classifiers (see Table 4.6). As a secondary search, we implemented similarity search. This allowed users to query using exemplar images as search criteria, see Figure 4.3. We represent each video based on the number of shots, with random selection of keyframes of videos containing more than 10 keyframes.


Carrier
12:58 PM
100%
Topic: 0302
Time Remaining: 4:09
growing in Ohio, Game Show fur cats, and Chi. . .
(21/100)
openhagen; Yet it could be anywhere in Denma. . .
(22/100)
Ela bebe um pouco de água do seu potinho
Catraca gosta de ficar embaixo da janela
n dog s life. a day in the life of a very laz. . .
(23/100)
my blog, but have cross-posted here, as we. . .
(24/100)
Reset Concepts
Search

boat

OFF	Animal
OFF	Beach
OFF	Beard
OFF	Black & White
ON	Boat/Ship
OFF	Building
OFF	Bus
OFF	Car
OFF	Charts
OFF	Cityscape
OFF	Computers
OFF	Computer Screen
OFF	Crowd
OFF	Daytime Outdoor
OFF	Face
OFF	Flower
OFF	Ground Vehicle

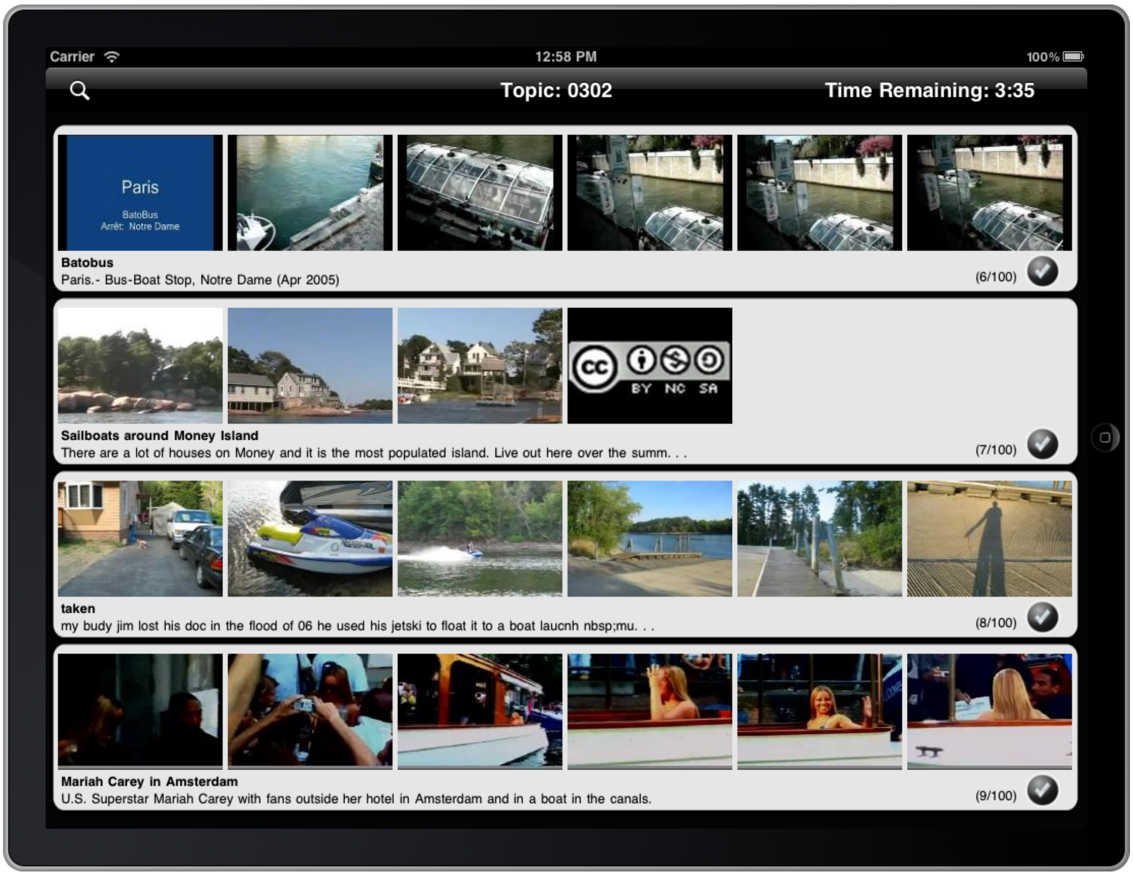
Figure 4.1: An example user search interaction with the Known-Item Search 2010 system

We recruited the 16 participants outlined in the user section. Each participant was provided with a list of written instructions which included a list of topics and an assortment of user surveys to be carried out. Users were required to capture topic level and overall experiment feedback. We also captured information regarding demographic profile and familiarity towards this experiment.

Each participant in the test was informed that they would be performing a search task on a Content-Based Information Retrieval system. We gave the users

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

Table 4.7: Table outlining the topic distribution between the novice and expert user groups



Batobus
Paris.- Bus-Boat Stop, Notre Dame (Apr 2005) (6/100)

Sailboats around Money Island
There are a lot of houses on Money and it is the most populated island. Live out here over the summ. . . (7/100)

taken
my budy jim lost his doc in the flood of 06 he used his jetski to float it to a boat laucnh nbsp;mu. . . (8/100)

Mariah Carey in Amsterdam
U.S. Superstar Mariah Carey with fans outside her hotel in Amsterdam and in a boat in the canals. (9/100)

Figure 4.2: An example result from interaction with the Known-Item Search 2010 system

a brief demonstration of the system by running through one full training topic, showing the users all of the functionality and features, finishing with successful identification of the video relating to the topic description. We decided to run the experimentation in groups of two, allowing for easier monitoring of participants. The users were informed that they would be required to search through the collection to find a single ”known-item” per topic. Each user was assigned 12 topics, allocated based on a Latin squares design (see Table 4.6). This layout of

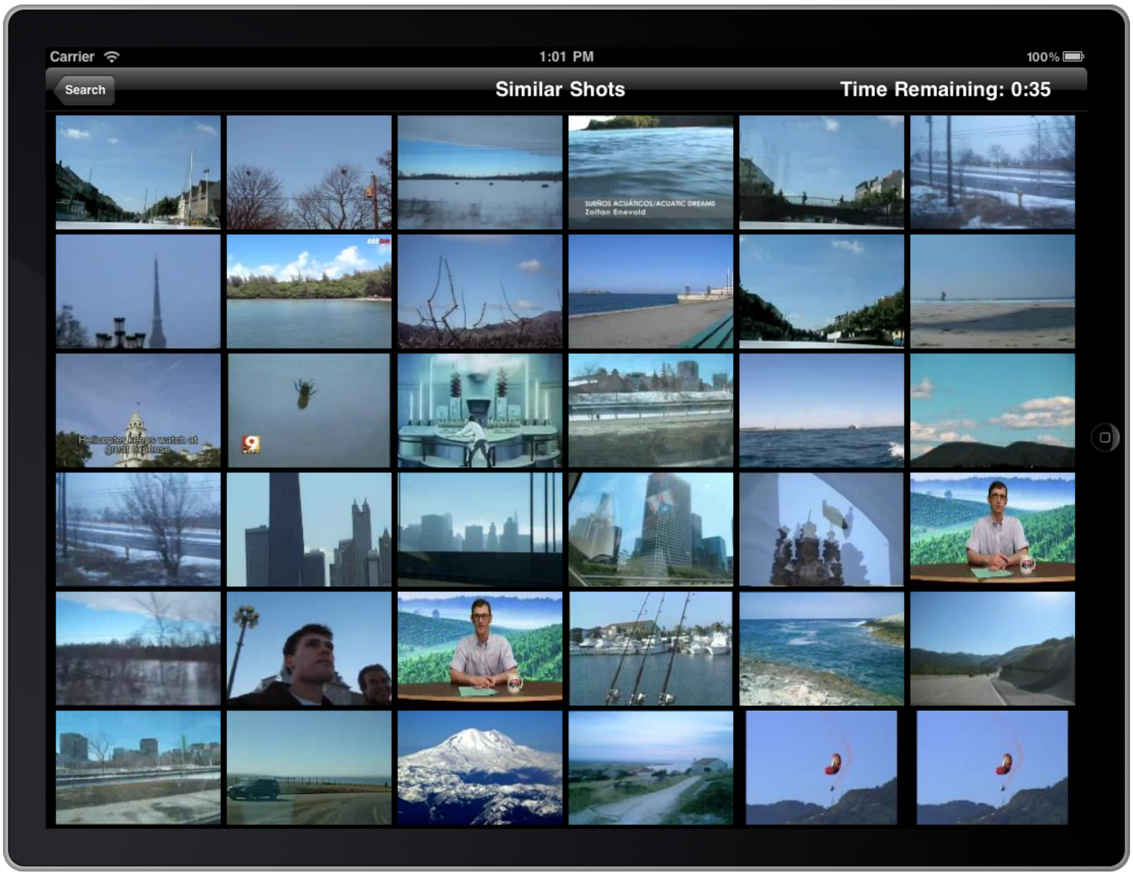


Figure 4.3: An example similarity search result from Known-Item Search 2010 system

topics ensured that each topic had equal coverage and that the topics were evenly distributed throughout the users by random ordering. Due to the users receiving training topics we do not believe that the random ordering will have any effect on the search performance. Before beginning the TRECVid experiments, we allow the users 30 minutes hands on experience with an iPad and 30 training topics to allow them to familiarize themselves with the system. Once confident, the user confirmed that they were ready to go and comfortable with the training received before proceeding. Users were then asked to carry out the experiments topic by topic, pausing after each topic to give feedback. All of our participants finished their topics in a single session.

4.5.1 Results

Figure 4.4 presents the results for all submitting teams with respect to Mean Elapsed Time until the known-item is found. Our two runs are highlighted. Figure 4.5 presents the results based on Mean Inverted Rank, a measure for the number of items found from the assigned topics. Both runs represent results from multiple users where we have picked the best time for each topic in order to populate our submission.

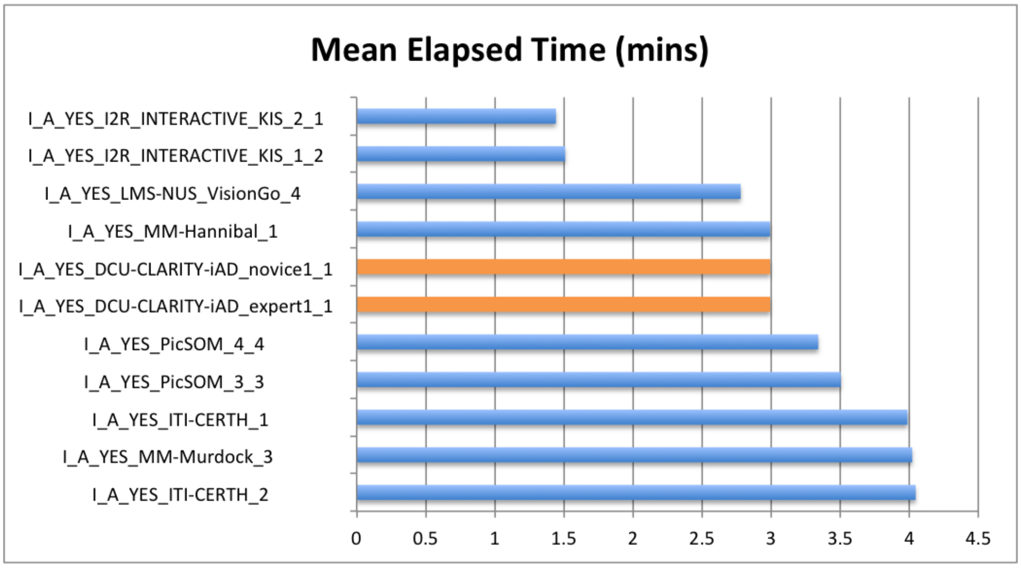


Figure 4.4: TRECVid Mean Elapsed Time results for the 11 participant runs, our submission highlighted in orange (lower is better)

The interactive known-item search task at TRECVid 2010 had 6 teams submit a total of 11 runs. There was participation from veteran groups such as the University of Amsterdam Media Mill team and Carnegie Mellon Infomedia team

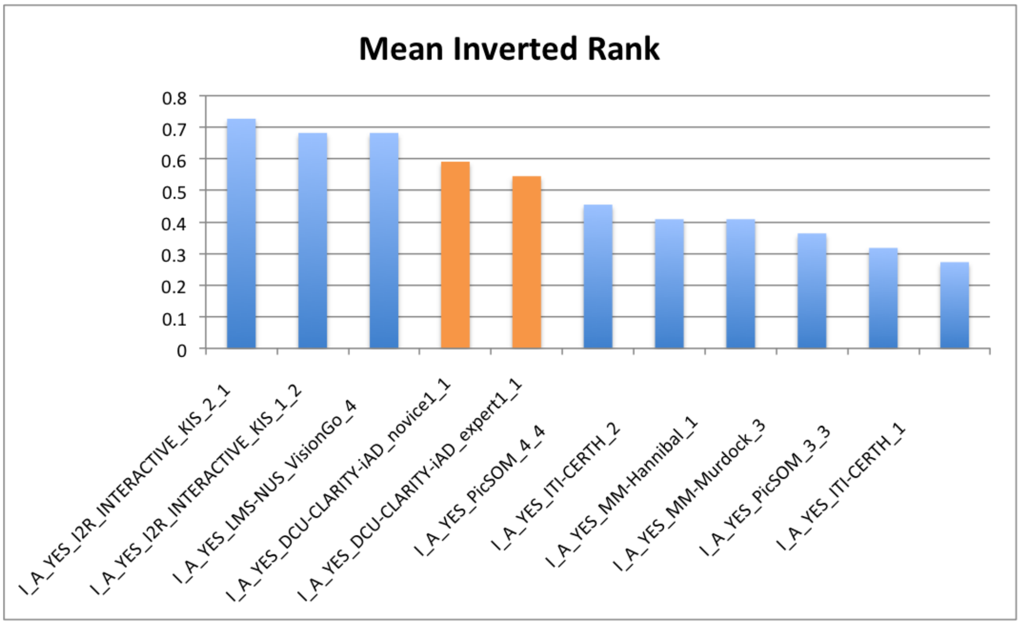


Figure 4.5: TRECVid Mean Inverted Rank results for the 11 participant runs, our submission highlighted in orange (higher is better)

who have participated since the beginning of TRECVid and are renowned experts in the field of video retrieval.

The 24 topics initially set to form the queries for our system were reduced by 2 due to multiple near duplicates in the collection. We also discovered from the experiments that the topics fell into one of two categories of topic, Hard or Easy. Easy topics allowed users to find the target document in few searches and with a minimum amount of time spent on topic. Conversely, hard topics took a large number of creative searches to be performed and a long period of time to discover the item. Our users, both expert and novice, found all of the assigned easy topics but few of the hard topics. Upon further investigation of the known-items it was found the majority of the hard topics had little or no meta-data associated with them and as such were relying solely on the visual elements of our search system to locate them. We will further investigate hard topics in our discussion section.

Overall our runs were ranked 5th and 6th based on the mean elapsed time and 4th and 5th based on Mean Inverted Rank. In the expert run there were a total of

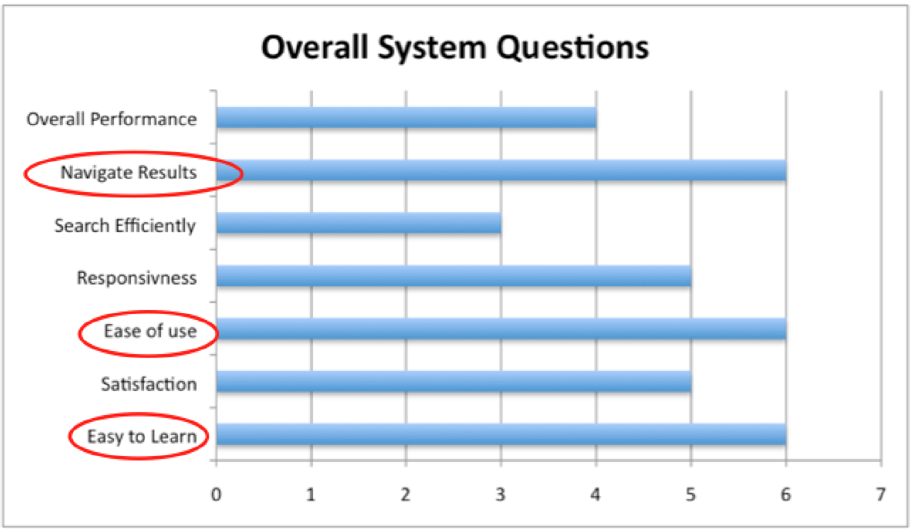
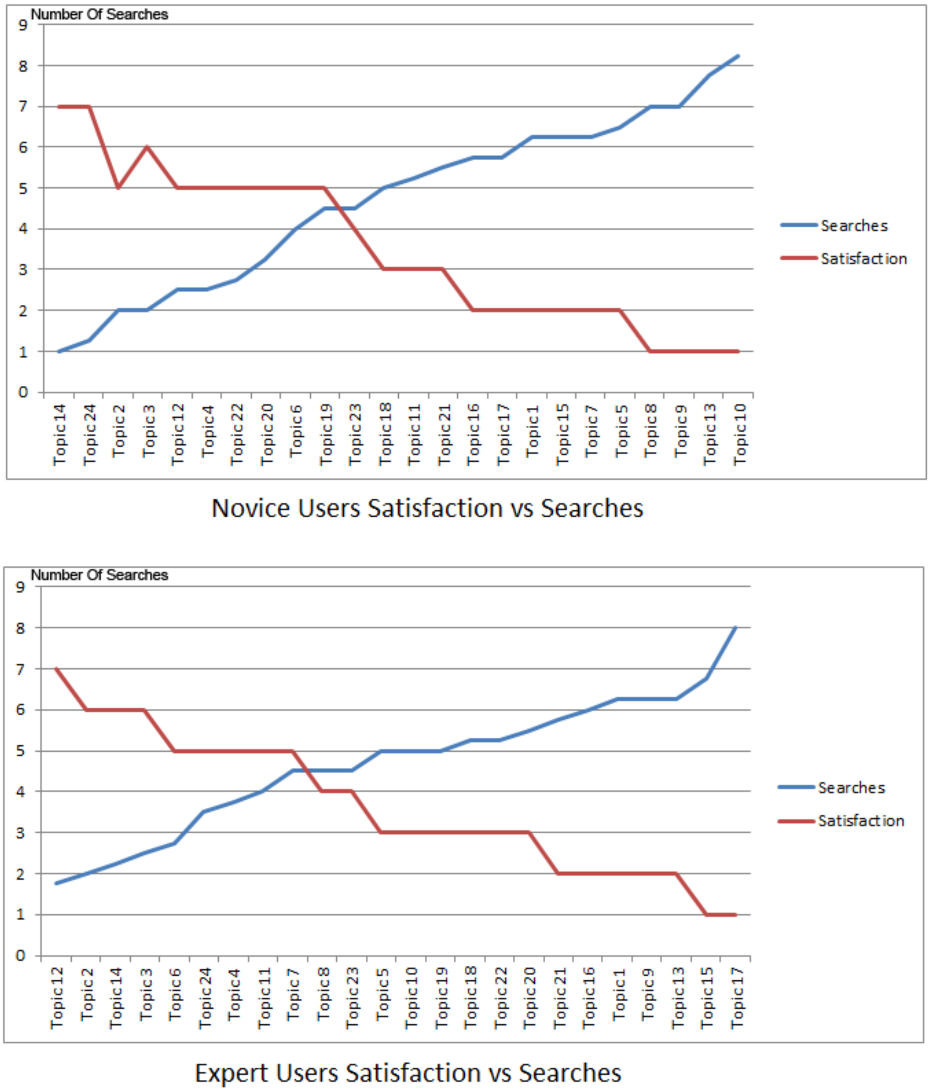


Figure 4.6: Overall system questionnaire results

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 these topics is not surprising. Having observed the user experiments, it was clear that users found the topics to be very easy or very difficult. Perhaps more interestingly, 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. While our overall user satisfaction for the system was quite high, users found it quite easy to use in terms of navigating results, ease of use and ease at learning (Figure 4.6), when we view our topic level feedback we see that user satisfaction in this case is directly related to topic difficulty. Easier topics which the user subsequently found the known item are among the highest scoring in terms of satisfaction, whereas difficult topics which cause the user to use multiple searches ending in the know-item not being found had the lowest satisfaction score. We see this visualized in Figure 4.7.



Novice Users Satisfaction vs Searches

Expert Users Satisfaction vs Searches

Figure 4.7: Mapping user satisfaction to topic searches

4.6 Discussion

In this section, we begin by discussing the results of the user study, paying particular attention to the captured system logs, next we analyze the classifiers to judge the best performing, then explore the feedback attained through the user surveys.

	Novice	**Expert**
Topic 1	6.25	6.25
Topic 2	2	2
Topic 3	2	2.5
Topic 4	2.5	3.75
Topic 5	6.5	5
Topic 6	4	2.75
Topic 7	6.25	4.5
Topic 8	7	4.5
Topic 9	7	6.25
Topic 10	8.25	5
Topic 11	5.25	4
Topic 12	2.5	1.75
Topic 13	7.75	6.25
Topic 14	1	2.25
Topic 15	6.25	6.75
Topic 16	5.75	6
Topic 17	5.75	8
Topic 18	5	5.25
Topic 19	4.5	5
Topic 20	3.25	5.5
Topic 21	5.5	5.75
Topic 22	2.75	5.25
Topic 23	4.5	4.5
Topic 24	1.25	3.5
Average:	**4.70**	**4.68**
Standard Deviation:	**2.11**	**1.62**
F-measure P:		**0.21**

Table 4.8: Topic by topic average number of searches by each user group, with the average number of searches per run almost identical

4.6.1 Result Discussion

Through user testing we wanted to compare the performance of our user groups. In particular we wanted to see, if by developing a targeted content-based video search system, we could bridge the gap in expertise between novice and expert users. From the official results evaluated by NIST, we can see that both user groups perform similarly with respect to mean elapsed time, both groups just shy of the three minute mark per topic. We also see, from the official results, that

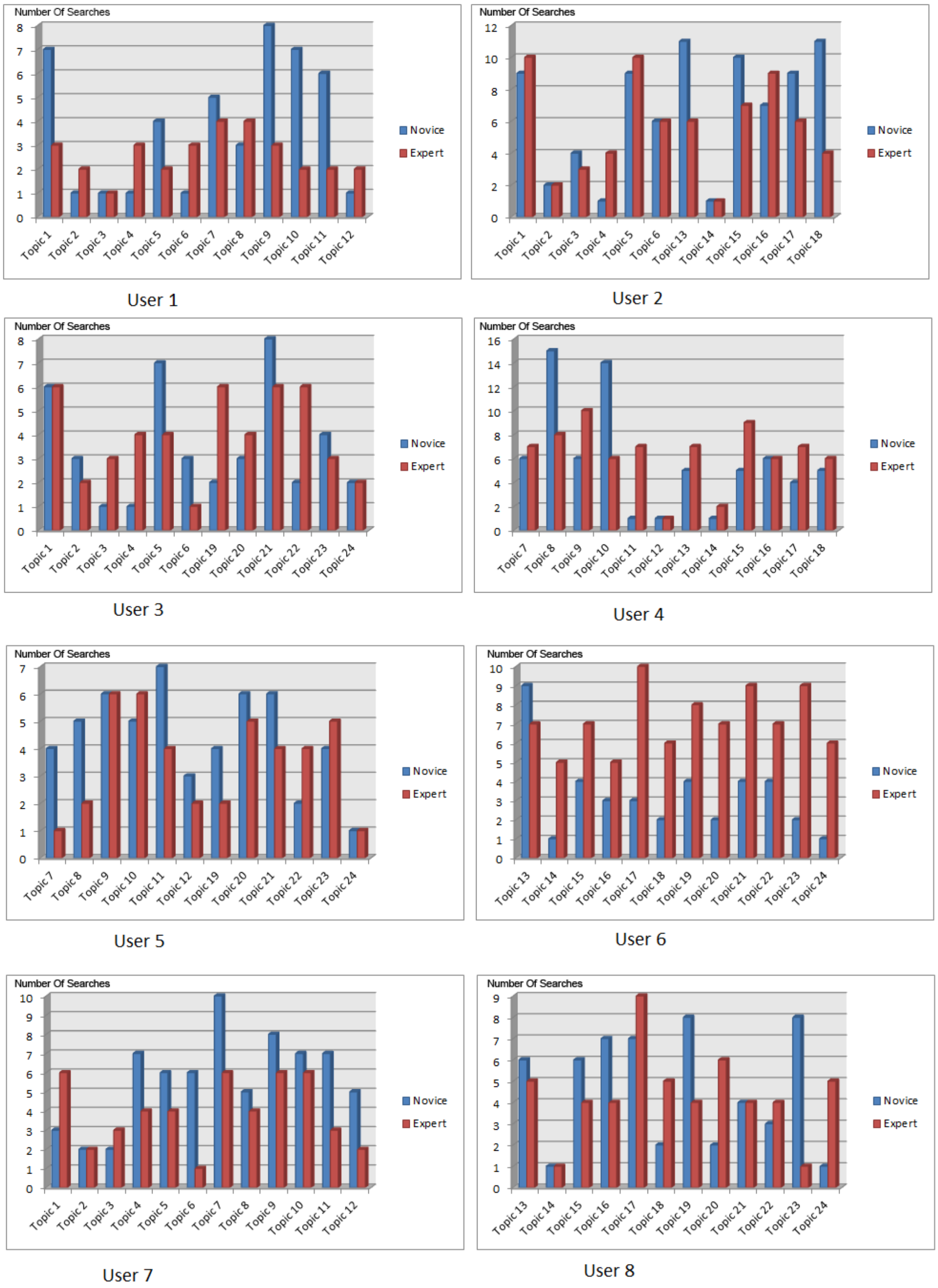


Figure 4.8: An overview of searches performed by both our expert and novice user groups with direct comparison on the same assigned topics

both user groups found a similar number of known items in this test, though our novice users found an extra item. Further to this we analyzed the users interactions with the system. We looked to the usability captured from our system logs (see Appendix C). In Figure 4.8 we see a visualization of our eight expert and

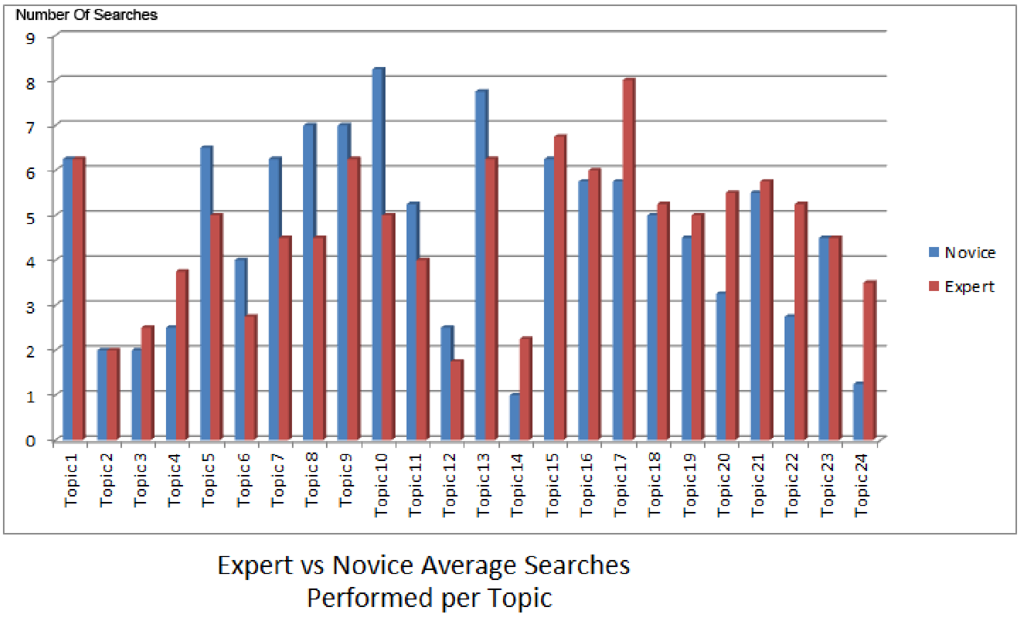


Figure 4.9: Graph depicting the average of participants 1 - 8 of both expert and novice user groups, shows search strategy more inline with each other

novice user tests side by side with respect to the amount of searches performed per topic. On visual inspection of the graphs, we see that in the case of User 1 our users share similarity on 5 topics, with the expert users clearly better on 4 separate topics. In the case of User 2 we again see 5 of the known items getting similar numbers in search with the experts performing better than the novices on 4 other topics. User 3 we again see similarity in 5 of the assigned topics this time. However, the novice users outperformed the experts in 4 of the assigned topics by a noticeable margin. For User 4 we notice seven of the assigned topics share similarity with the experts clearly beating the novices in 2 topics. In the case of User 5 we see again five similar topics with the experts marginally better in a further 3. In the case of User 6 we see similarity in only a single case, with the novice user performing better in the other 11 cases, we believe this to be the case due to the test being performed by the most experienced novice user against the least experienced expert user, a random event. The vice versa is true of User 7 where our most experienced expert user gains better performance in 9 topics with similarity in only one case. Finally the User 8 group exhibits the most close similarity with half of the topics being similar with a share of 3 and 3 for the remaining topics between the novices and experts.

From these results we get a very mixed view, in some cases our experts perform better and in others our novices, taking both groups of eight users averages on each of the 12 assigned topics we see in Figure 4.10 that the graph shows much more similarity between the two, with our average expert and novice user only slightly better on 4 topics each, sharing similarity on 20 topics. From Table 4.8, we see that taking an average of all searches both the experts and novice perform approximately 4.7 searches each. We can assume by measuring the Standard Deviation and then the F-Measure that from the results attained in this experiment our user groups are not significantly different as the P value is not less than 0.05. From this we can conclude that both sets of users perform similarly.

4.6.2 Further Log Analysis

We see a visualization of the search approaches performed by each of our user groups in Figure 4.10. From the graph, we can see that our novice users rely more on text based searches, with almost 50% of searches relying solely on text. In total 90% of the searches contain text in some form, with less than 10% of cases featuring only searches based on visual features. Our experts are more accepting of visual features, with more than 20% of cases featuring a search which uses a non-textual query element. We believe with our novice users being more familiar with YouTube type system they are more comfortable with only using text based searches, we believe that in future systems we should aim to incorporate these visual features automatically. This would allow us to combat situations where users are presented with "Hard Topics" and text search is of little help, we will investigate this further in Chapter 5, where we will address our second hypothesis.

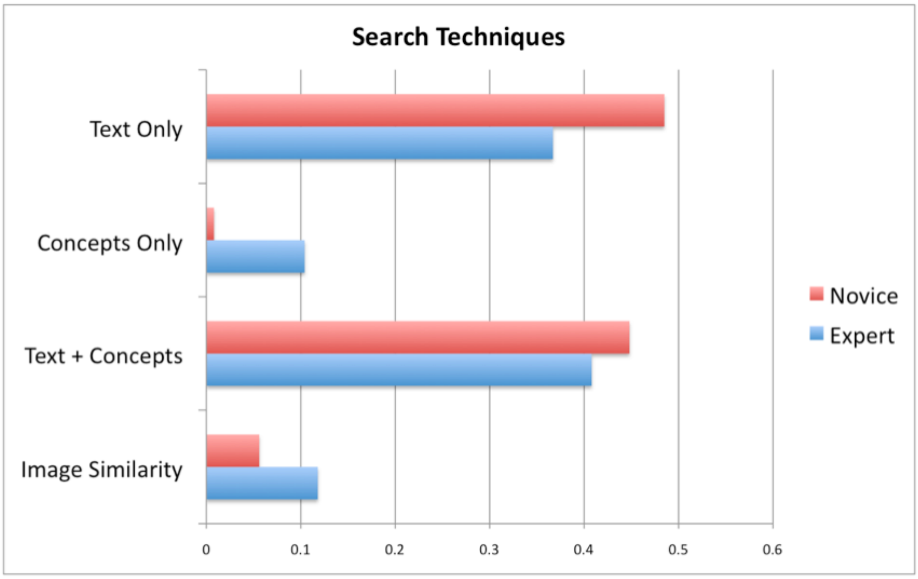


Figure 4.10: An overview of the search techniques carried out by both the novice and expert users

4.6.3 User Feedback

We capture users' feedback both implicitly and explicitly. Our implicit method relies on the capture of users' search behaviors and is attained by logging the users' interactions with the search service. We also capture user data explicitly through the use of forms, which the users are asked to fill in during the experimentation process. These forms can be see in Appendix A.

Visual Features

As we see from the previous section, our novice users were quite reluctant to utilize both the classifiers and similarity search that they had available to them. Even those that used the classifiers for search focused on a select few, ignoring the vast majority of the 33 trained classifiers. We see in Table 4.9 that only eight

Usage	Concepts
> 50	animal, person, indoor, car, vegetation, landscape, building, bus, cityscape
49 - 15	boat/ship, computer screen, crowd, face, ground vehicle, military, outdoor, tree
< 15	meeting, nighttime, road, sky, office, beard, computers, flower
unused	Black and White video, daytime outdoor, indoor sports, map, charts, beach, stadium, snow

Table 4.9: Usage of classifiers based on overall search in user testing

classifiers out of the 33 are used over 50 times each, with classifiers such as person and indoor being utilized most commonly. From our trained concepts, we notice that nine concepts are not used at all, with a further eight being used less than 15 times each. Our experts were more inclined to utilize the visual features with 63% of first search attempts featuring one or more of the provided classifiers. Finally, in 10% of cases an expert user performed a similarity search.

From the retrieved feedback, the novice users informed us that using the classifiers was quite daunting. Most believed that certain classifiers were unusable. In some cases the classifiers were not applicable to the query and in others, as shown from the evaluation in the Semantic Indexing Task (see Chapter 3, Task Description), the classifiers' performance was quite low. For future systems we must reduce the number of classifiers to a more manageable amount, and carefully select based on examining those which were used most often in this experiment. Our users were not satisfied with low performing classifiers which added nothing except noise to the returned search results. By utilizing only the classifiers which exhibit high confidences based on the training data, we believe users will be more likely to explicitly select a classifier for search. We will investigate this further in Chapter 5.

Hard Topics

One major frustration point for our users was the fact that the known-item for certain topics could not be found. In fact, there were 6 topics which were not found by any of the participating groups to TRECVid. These topics later coined "Hard" topics had little or no meta-data associated with them. Three of these topics had only title information associated with the meta-data and no description text or associated keywords. The other topics featured single unrelated keywords and sparse, one or two word, descriptions. As such, they were very difficult to find, especially for our novice group, who as we had determined earlier relied heavily on text search.

There were 10 topics in the collection which satisfied the criteria of Hard topics. Of those only two were found by our system both of which were found by our novice users and only one of which by our expert. Each of these queries had required the use of visual features to find the item, with respect to the hard item that both the novices and experts found, the experts performed less searches due to the uptake of visual features earlier in their search strategy. We will discuss Hard topics more in the succeeding chapters.

4.6.4 Research Questions

For the research carried out in this chapter, we proposed a number of research questions:

1. Will using a tailored interface design on handheld devices impact performance when compared to other state-of-the-art systems participating in video bench marking conferences?

We see from the results section, with particular focus on the TRECVid metrics of Mean Elapsed Time and Mean Inverted Rank, that our system performs at a median level when compared with the other participating research groups. Given that our system did not under perform compared to these other systems, we can determine that the tailored interface and handheld device does not impact on the performance of the system.

2. What visual features will allow our inexperienced users to take advantage of content based search? How will our users interact with the features? How frequently are the features used?

From post experiment analysis we looked at methods of visual search with which our participants used most. We found that of the visual aids, similarity search was used rarely. Concept search, though used more often, only seven of the available classifiers were used frequently.

4.7 Conclusion

In this chapter, through both automatic and user testing means, we provided a framework which satisfies the criteria outlined in our first hypothesis:

> *Using a tailored interface design, which utilizes selected content based retrieval techniques on handheld devices will increase the performance of novice users when carrying out known-item search tasks.*

We have shown that by developing a simple intuitive iPad interface which utilized a set of targeted content base techniques, we can achieve similar results from both an expert and novice user group. Our official runs for the interactive known item search task show that our experts and novices perform alike with regard to the relevance measure Mean Elapsed Time. Furthermore, we have shown the the search strategy adapted across the eight novice and expert user groups, have shown similarity with respect to the number of searches carried out on average. When compared to the greater research community who also submitted to TRECVid our proposed intuitive interface performs to a better than median level.

The results attained through this research warrant further investigation, while the system did perform well with regard to the outlined experiments. We believe that further optimization is required to further enhance the power of the novice user. This work features only basic visual features which user must explicitly supply to formulate queries. One of the shortcomings of this is by putting the onus on the user to supply the criteria of search. As we see from the results, users are not comfortable with interacting with the system in this way.

Through further analysis of interaction logs, we determined points of improvement with which to implement in future systems. The understanding from this experiment is to seamlessly integrate visual search in a way which is transparent to the underlying user. We believe this method will greatly improve the overall performance of the system. We intend to evaluate this approach in the next chapter.

Chapter 5

The Cluster List

5.1 Introduction

In the last chapter, we presented our pilot system with which we participated in the TRECVid 2010 Known-Item search task. This system featured a selection of both visual and textual search methodologies which had been configured to aid the user with respect to finding the assigned topics. Through experimentation we evaluated the performance of both a Novice and Expert user group, and while both groups performed similarly, it led to a number of questions regarding the search strategies, specifically those of the novice users. It was found that novice users were reluctant to adapt visual search despite post experiment cases showing improved rank of the known-item. Upon further analysis of both logs and user feedback, it was discovered that users found the visual search elements too complex, not knowing where and when to issue an appropriate query. Other users lost confidence in the classification early on in the experiment and relied solely on text for the remainder. We believed that we must remove the complexity of this visual search by seamlessly integrating visual search with user-generated queries. In doing so we can aid the user in finding the known items. This has led to the creation of our second hypothesis:

Taking a single keyframe representation approach, where the keyframe is identified by content-based techniques, grouping similar items will help a user to more quickly locate/dismiss relevant videos.

In the case of the known-item search task, we assume that by using a clustering technique we can emulate the similarity search of the previous system without the user overhead. Also by precomputing a single keyframe representation we will also effectively remove the complexity of selecting appropriate classifiers. To test the effectiveness of this hypothesis we formulated the following research questions:

- How do we best display to the user an accurate video representation? Is a single keyframe sufficient? Can the automatic use of classifiers help with this representation?
- By grouping content can we provide users with a better search experience? Does a ranked list of clusters perform better than a standard ranked list?

We began by using clustering to determine a keyframe representation which will increase the likelihood of populating the entire interface canvas. Our next task is concerned with retrieving the correct representation of the content, we experiment with different types of single keyframe representation methodologies, matching them to both the training topics and videos based on a rating assigned through user testing. From here and based on the novice users lack of usage of similarity search in the previous chapter, we employed a basic clustering algorithm to group like content. We aimed to test if re-ranking based on visual techniques would aid our novice users further. We evaluated both experimental sets using the TRECVid 2011 evaluation framework with which we are provided with generic data from the Internet Archive Creative Commons (iacc.1.b).

With regard to user testing, we again employ novice users, casual internet users with no formal knowledge of visual search. We use different users for each experiment, with the configuration stage falling to non-technical summer interns and the TRECVid testing to Norwegian Business School students. Both sets of users satisfy the criteria of novice users outlined in chapter 4.

In the remainder of this chapter, we start by discussing the data which we incorporate into our prototype system. Next we outline the two sets of novice users we recruit for both of our user experiments. Then we evaluate our automatic experiments used in the configuration of the system. After this we expand upon our user experiments, testing our representation method and approach towards clustering. Finally we close this chapter with a discussion of findings and a brief conclusion.

5.2 Data

The experiments outlined in this chapter utilized the second iteration of the Internet Archive Creative Commons (iacc.1.b) from the TRECVid 2011 data corpus. Given that this data is from the same collection as that presented in chapter 4, it is consistent and we are provided with similar textual data such as meta-data from author annotated video details and Automatic Speech Recognition extracted through a machine learning phase, leading to a transcript of the spoken word (see chapter 3 for description). We are also provided with visual ground-truth from a collaborative annotation task run by NIST to aid in the design of our classification models.

The data itself is further split into training and test sets, both datasets contain generic internet video which can encompass both professionally edited and User Generated Content (UGC). We are provided with over 200 hours of test video up to 3.5 minutes in duration. The training data is adapted from the previous year's

Structure of the Provided Data TRECVid 2011

Hours Of Video in Test Collection	200 (approx.)
Number of Videos in the Training Collection	8471
Number of Shots in the Training Collection	144935
Number of Videos in the Test Collection	8216
Number of Shots in the Test Collection	137327
Number of Topics in the Training Collection	300
Number of Topics in the Test Collection	24

Table 5.1: Overview of the data used for TRECVid 2011

test data and has undergone much post-experiment analysis as seen in Chapter 4. In Table 5.1 we see a breakdown of the data used in the experiments for TRECVid 2011 with the training and test sets featuring similar numbers of both shots and videos. A more detailed description of TRECVid style datasets can be found in Chapter 4 Section 4.2.

5.2.1 Additional Data Sources

Three additional sources of data were required to aid in the design of the visual search components, they were:

- Classifier ranked-list
- Video representation list
- Cluster list

We extracted visual features in the form of MPEG-7 (Edge, Scalable Colour and Colour Histogram) and OpponentSIFT (van de Sande et al., 2011) from each of the keyframes attained through the shot boundary master file. OpponentSIFT extends Scale Invariant Feature Transformation (SIFT) by utilizing color descriptors to extract interest points which provide a feature description of the content. These ″low-level″ features aided in the generation of the classifiers. Similar to chapter 4, we use an SVM to train the classifiers based on positive and negative examples attained through the visual ground-truth. A ranked list is generated for each of the trained classifiers based on the probability of containing the concept, these lists form the basis of the classification search in the final system.

The extracted visual features are further used to build a video representation list, which in turn is used to identify the most appropriate keyframe representation for a given video. When a user has provided only a text based query, the most relevant keyframe from a classifier which has scored over a certain confidence threshold is used, further explained in Section 5.4.

Lastly, we used the visual features to determine similarity of the video representation by employing a visual clustering algorithm. In this way, we provided a ranked list of documents where those without meta-data could be boosted by association through visual similarity with an item containing meta-data. Section 5.4 and 5.5 have more information on the clustering method employed.

5.3 Users

In the previous chapter, we ran experiments with novice and expert groups. In this and subsequent chapters, we focus on the novice users and will solely use a group with limited experience in content based search to test our theories. In this chapter, we have proposed two user based experiments, the first experiment tests the user's preference towards the relationship of both keyframe-topic and keyframe video. The second experiment requires the users to perform an experiment on a retrieval system over the TRECVid corpus for 2011. We required two separate groups for our testing. Table 5.2 outlines the demographics of the groups.

- **Experiment 1**: For this experiment, ten student interns based in DCU for the summer months volunteered to run the experiment. These interns had not yet achieved an undergraduate qualification, the majority having just finished third year in a science based course. The male candidates were students from computer science/business backgrounds, with no involvement in Information Retrieval. The two female users both came from a biology background, having limited computer experience outside of casual personal use. When asked how they relied on search system such as Google/YouTube younger members of the test group used these system more than the older members. Also the men tended to be more willing to source information online. The average age of the participants was 23.8 with a standard deviation of 2.57. These users represent our target novice user group for our first experiment in this chapter.

- **Experiment 2**: For this experiment we again recruited eight participants from one of our collaborative partners, the Norwegian Business School, Oslo. These users represented a mix of both undergraduate and post-graduate students. A faculty member responsible for the student groups also ran a user test. These users perform frequent searches on both YouTube and Google, but have limited knowledge of search systems outside of this scope. Six of the users had handheld smart phones such as the iPhone, with two from this group also possessing a tablet PC. The users ranged in age from 22 - 50, we noted that younger users were more likely to own a smart device with all users under 28 having a smart phone. The users over 35 were more likely to have both smart phone and tablet PC. The average age of the participants was 26.75 with a standard deviation of 3.28. These users represent an ideal novice group to test our system developed for our second experiment in this chapter.

Participant Profile		Experiment 1	Experiment 2
Age:	18 - 25	6	2
	25 - 30	4	5
	30 - 35	0	1
Web Search (inc Video)	Regular	7	6
	Infrequent	3	2
Handheld Usage	Never	5	1
	Infrequent	3	3
	Regular	2	4
Education	Undergraduate/No Degree	10	5
	Graduate	0	2
	Faculty/Staff	0	1
Gender	Male	8	2
	Female	2	6

Table 5.2: Participants' Profile for User Experiment 1 & 2

5.4 System Configuration Phase

Before our final version system for TRECVid 2011, we examined two different methods of representing video content to aid in evaluating our hypothesis and provide the users with a better search experience. The first of these configuration experiments tested a clustering algorithm, to identify a method of maximizing the use of the available interface canvas. The second experiment deals with matching content, both topic description and video, to a representative keyframe based on different frame selection methods.

5.4.1 Evaluation of Clustering

For this small heuristic experiment we wished to identify a method which would provide semantic grouping of content, while also maximizing the usage of the available screen. To achieve this, we implemented a the k-means clustering algorithm, a popular method, which allowed us to have full control over the number of keyframe groups. Using a keyframe layout set by Hyowon Lee in previous TRECVids (Foley et al., 2005; O'Connor et al., 2006), our interface can

		k = 50	**k = 100**	**k = 500**
	Returned	per Cluster	Per Cluster	Per Cluster
Topic 1	283	8.84 (32)	4.49 (63)	1.41 (201)
Topic 2	84	5.6 (15)	2.9 (29)	1.38 (61)
Topic 3	462	12.83 (36)	5.78 (80)	1.2 (386)
Topic 4	176	4.09 (43)	2.63 (67)	1.49 (118)
Topic 5	90	4.5 (20)	1.67 (54)	1.14 (79)
Average	219	7.17 (29.2)	3.49 (58.6)	1.32 (169)

Table 5.3: Table outlining the chosen number of items per cluster based on training topics using different values for k (Number of Clusters in Brackets)

comfortably accommodate five keyframes per line. We required our keyframe groups to be populated with a similar number to take full advantage of the display.

We used data from the training set detailed above to facilitate our experiments. Feature extraction in the form of MPEG-7 edge, color histogram and color layout provided the low level features. We ran the clustering algorithm three times using the extracted features with differing *k* values of 50, 100 and 500. We built a prototype retrieval system over the training data using text only search, to simulate the novice user's preferred primary search (from Chapter 4). We incorporated the clustering visualization to this prototype system and issued five training topics as a test to determine which value of *k* allows for maximized representation.

Results

We see from Table 5.3, that utilizing a value of *k* equal to 100 will net an average number of frames per cluster of 3.5 which falls within the range we want for our representative clusters. Using the smaller cluster size while also viable, will lead to sideways scrolling which we tried to avoid, as users tended not to avail of it in previous experiments (TV2010). We found that using larger cluster sizes will lead to much more blank canvas with an average of only greater than one frame per cluster.

5.4.2 Optimizing Visual Representation

Our previous system developed for TRECVid used visual search techniques in conjunction with interface design to determine baseline features to aid Novice user groups. While we posted a favourable score, post analysis revealed novice users had reluctantly availed of the provided visual features. While these features improved the rank in certain cases, it was decided that an automatic approach to applying these visual features would help. We proposed the following experiment to test this theory.

Selection of Representative Keyframe

In Chapter 4 we utilized a random sampling of keyframes to represent each video. This method caused not only keyframe duplicates, but certain cases omitted helpful frames. This was more often the case in videos with a large number of shots. In this experiment, we devised methods to limit the shortcomings of the previous system. Beginning by identifying single keyframe representation which accurately represent the content (Browne et al., 2000, 2001). As such, we devised five representations with which to test outlined below:

- **Method 1 : First Keyframe:** the first shot in the video, certain shots captured the title, though the majority of cases a blank/black screen.
- **Method 2 : Middle Keyframe:** a standard representation approach within the field of shot-boundary detection for representing a video.
- **Method 3 : Last Keyframe:** the final shot of the video, in most cases contained a black screen but in certain cases contained a capture of title credits.
- **Method 4 : Random Keyframe:** can represent any keyframe selected from within the video, selection is randomized per topic.

	1	2	3	4	5	6	7
Classifiers	Adult	Vehicle	Building	Animal	Landscape	Plant	Indoor

Table 5.4: Hierarchical structure of the classifiers used for keyframe representation, manual selection of concepts based on confidence scores and likelihood in aiding the user

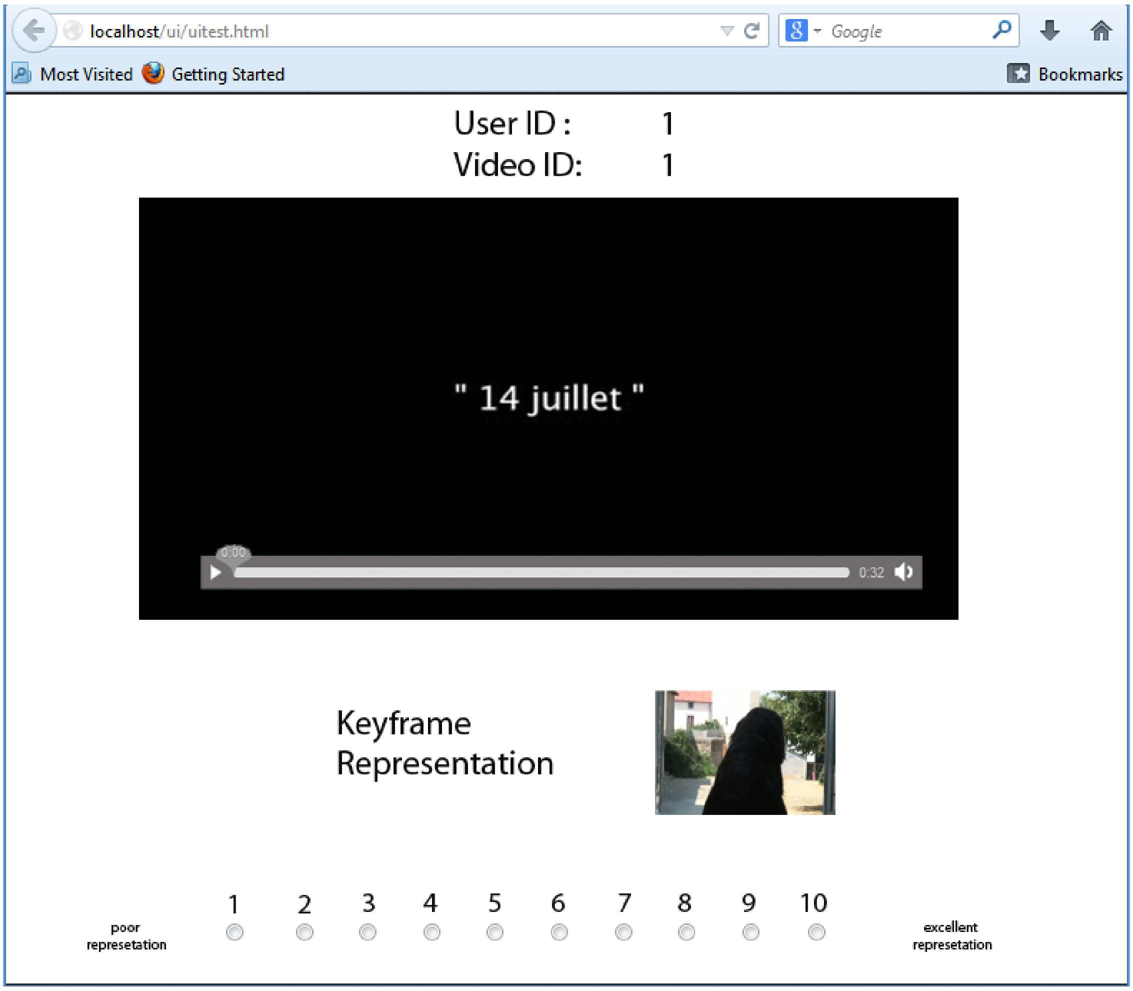
- **Method 5 : Visually Calculated.** Based on evidences found by the use of both features and classifiers to determine the best representation.

The visually calculated keyframe employed the use of the MPEG-7 features to determine a baseline representation. The average is calculated for each video based on the sum of the keyframes' MPEG-7 vectors. We use this average to determine the shot representation of the video with minimum distance from the average, based on the minimum distance formula. This keyframe will be used when no other influences, user or automatic, impact upon the representation.

$$MinDist = min \sum_{x=1}^{n} \sqrt{(x_1 - avg_1)^2 + (x_2 - avg_2)^2 + ... + (x_n - avg_n)^2} \qquad (5.1)$$

We took the set of classifiers trained using the positive and negative examples, which were supplied as part of the visual-groundtruth from the TRECVid framework, and applied them to each shot. We have determined a hierarchical structure to implement these classifiers automatically, outlined in Table 5.4. Keyframes which exhibit a high positive probability of containing a concept and are of higher importance in our above table, have the greatest chance of representing their respective video.

In order to test our users' preference towards these keyframe representations, we developed a simple web interface outlined in Figure 5.1. Users were presented initially with a login screen where they input their user identification number. From here users were prompted to either proceed with the experiment, or take



User ID : 1

Video ID: 1

Keyframe Representation

poor represetation 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10 ○ excellent represetation

Figure 5.1: Experimental interface for keyframe representation experiment

	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
User 1	1	2	3	4	5
User 2	2	3	4	5	1
User 3	3	4	5	1	2
User 4	4	5	1	2	3
User 5	5	1	2	3	4
User 6	5	1	2	3	4
User 7	3	4	5	1	2
User 8	1	2	3	4	5
User 9	2	3	4	5	1
User 10	4	5	1	2	3

Table 5.5: Distribution of topics/videos through our user groups. 1 = First Keyframe, 2 = Middle keyframe, 3 = Last keyframe, 4 = Randomly Selected keyframe, 5 = Visually Calculated keyframe

a training topic to familiarize themselves with the system. After each training topic, users were prompted again to see if they were ready to begin the test, with fifteen training topics in total. Once the users were satisfied, the test began. They were assigned five topics with different keyframe representations for each topic. The distribution of the representations is based on a 5 x 5 Latin squares design with two pass representation per topic over ten users (see Table 5.5). Users are prompted on each topic to rate the keyframe representation on a scale from 1 (poor representation) to 10 (excellent representation). We define our best keyframe representation based on this rating.

Matching Keyframe to Topic Description

With the testing system built, we devised two small user tests which lasted no longer then ten minutes each. The first test attempted to rate the selected keyframe representation to the topic description. We see in Table 5.6 the ratings given, based on each representation per user. We take these results, and in Table 5.7, we see the average of these representations based on the method and topic. These tables show us that both method 1 and 3 attain a very low ranking and should not be

	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
User 1	1	1	1	1	10
User 2	7	1	9	9	1
User 3	1	1	9	1	9
User 4	6	1	1	8	1
User 5	8	9	8	1	1
User 6	8	9	8	1	1
User 7	1	1	7	1	9
User 8	1	1	4	3	9
User 9	8	1	7	7	1
User 10	7	1	1	8	1

Table 5.6: User ratings based on keyframes matched to topic descriptions.

	Method 1	**Method 2**	**Method 3**	**Method 4**	**Method 5**
Topic 1	1	7.5	1	6.5	*8*
Topic 2	9	1	1	1	*1*
Topic 3	1	8	2.5	8	*8*
Topic 4	2	8	1	2	*8*
Topic 5	2	9	1	1	*9.5*
Average	2.6	6.7	1.3	3.7	**6.9**

Table 5.7: keyframe selection matched to topic description averages

used in representing the videos. Method 4, the random implementation while better is still not good enough with less than 50% of the maximum rank. Finally, method 2 and 5 show scores of 6.7 and 6.9 respectively. Both of these methods according to this user test show the highest potential in accurately representing the topic description.

Matching Keyframe to Video

In the second part of this experiment, users rated keyframe representations with respect to their related video document. We see in Table 5.8 the scores achieved based on this user experiment. In Table 5.9, we again see that method 1 and 3 are the worst performing in terms of approval rating, with method 4 being of mixed review and only achieving about half the approval rating compared with the best

	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
User 1	1	6	1	3	8
User 2	8	1	9	8	1
User 3	3	4	9	1	8
User 4	7	7	1	7	1
User 5	9	7	7	1	1
User 6	9	9	7	1	1
User 7	5	1	7	1	8
User 8	1	3	3	5	8
User 9	7	1	6	7	1
User 10	7	7	1	8	1

Table 5.8: User ratings based on keyframes matched to videos.

	Method 1	**Method 1**	**Method 3**	**Method 4**	**Method 5**
Topic 1	1	7.5	4	7	*9*
Topic 2	8	4.5	1	2.5	*7*
Topic 3	1	7	2	7.5	*8*
Topic 4	1	7.5	1	4	*7.5*
Topic 5	1	8	1	1	*8*
Average	2.4	6.9	1.8	4.4	**7.9**

Table 5.9: keyframe selection matched to video averages

performing method. Again both our middle and visually calculated keyframe score very highly, with our visually calculated slightly besting the central frame.

Results

From the results attained through this round of user testing we can safely assume that our visually calculated keyframe is the users most preferential in representing this type of video content. We ran a significance test on the User, Topic pairs from both the middle and visually calculated representations. Unfortunately for us and possibly due to the size of the user study we were unable to show any significance, using a single tailed T-test approach where $p < 0.05$ defines significance. Nonetheless, we did show significance compared to the random keyframe selection, see Table 5.10. We will, however, use the visually calculated keyframe representation

	T(Method 2, Method 5)	T(Method 4, Method 5)
Keyframe to Topic	0.12	0.016
keyframe to Video	0.5	0.0009

Table 5.10: Testing significance of method 5, Significant if p <0.05 (T-test)

as an input into our next experiment dealing with keyframe clustering to aid the users.

5.5 TRECVid 2011: Evaluating Clustering

For this set of experiments we looked again to our TRECVid prototype system outlined in chapter 3, implemented using the data described in this chapter with our eight Norwegian business school users as testers. Each participant was assigned a list of instructions which includes topics to be carried out and a survey form to be completed both during and after testing. The topics were distributed using a Latin-squares model as described in Table 5.11. In this way each topic had two evaluations per system.

One of the changes implemented in this work was the use of Solr, replacing our previous index which was based on Terrier[1] developed by the University of Glasgow. We had also considered employing the output from the phonetic encoding tool. However, post-experimentation last year showed that while this did increase recall, it actually decreased the average rank of the known items, so it was not included in this experiment.

Accurate keyframe selection is especially important given that our experiment was heavily focused on the video ranked result representation. We employed two types of keyframe selection criteria. Firstly, the visually calculated keyframe is chosen based on the previous experiment. Secondly, we employ a query-biased keyframe selection approach when the user has entered visual classifiers

[1] http://terrier.org/

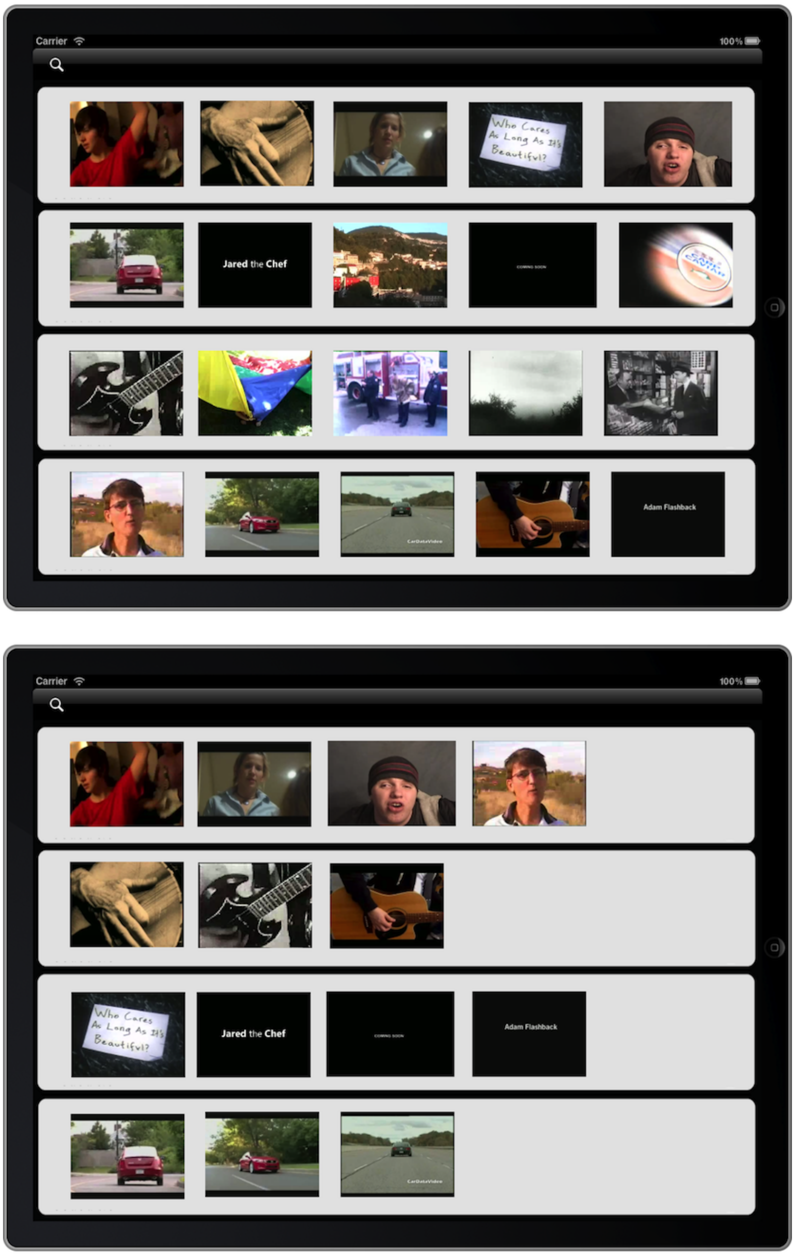


Figure 5.2: A view of the two interfaces used in this experiment. The top diagram shows the non-clustering system. In this system the results are disorganized, users must scrutinize each element as there is no common theme. Conversely the bottom diagram shows the interface utilizing clustering. In this system items are grouped based on similarity, in this way users can easily dismiss items at a group level, this leads to more efficient, speedier searches

to identify query-appropriate keyframes. For cases when a single visual classifier was selected, the top-ranking keyframe (for that classifier) is chosen. In the case where multiple classifiers were selected, evidence from all classifiers were fused to identify the top-ranked frame.

5.5.1 Results

Our two systems for comparison in 2011 were a single keyframe per video (WWW style) baseline system and a result clustering system, see Figure 5.2. The result clustering system allows for users to view items which exhibit similar features, those of the MPEG- 7 descriptors, and have them presented side-by-side. This allows the users to view visually similar content clustered together, and reduces the overhead of scrolling/browsing through the whole ranked list. We found that in most tasks, users found the known item faster on the clustering system, rather than the baseline approach.

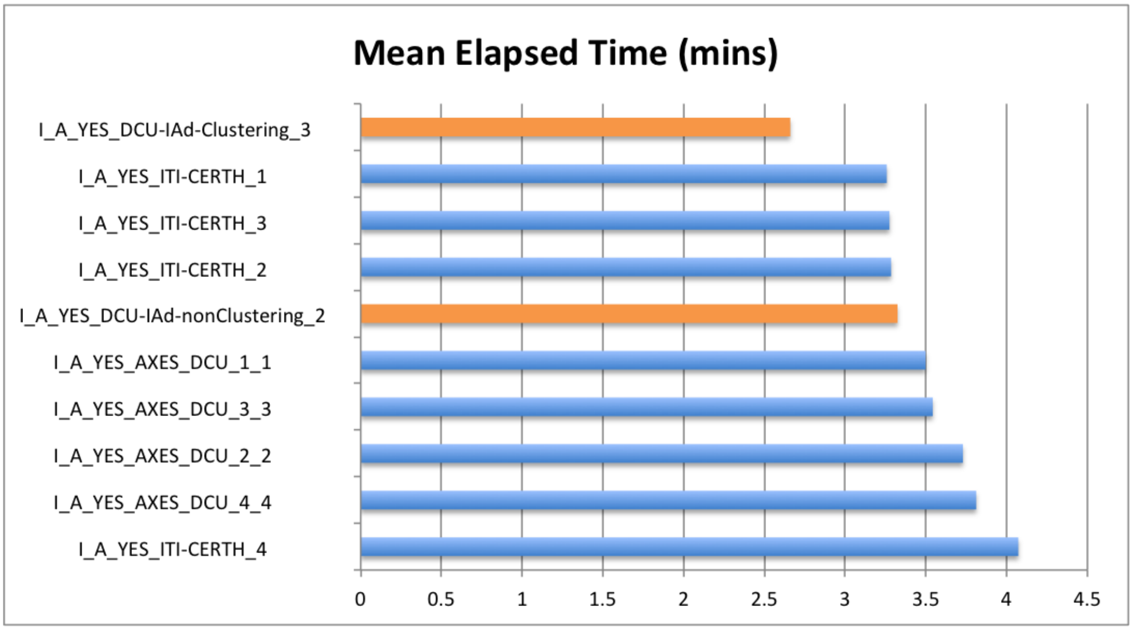


Figure 5.3: TRECVid Mean Elapsed Time results for the 10 participant runs, our submission highlighted in orange

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

Table 5.11: Table outlining the topic distribution over our eight participating users, (a) denotes the non-clustering system (b) denotes the clustering system

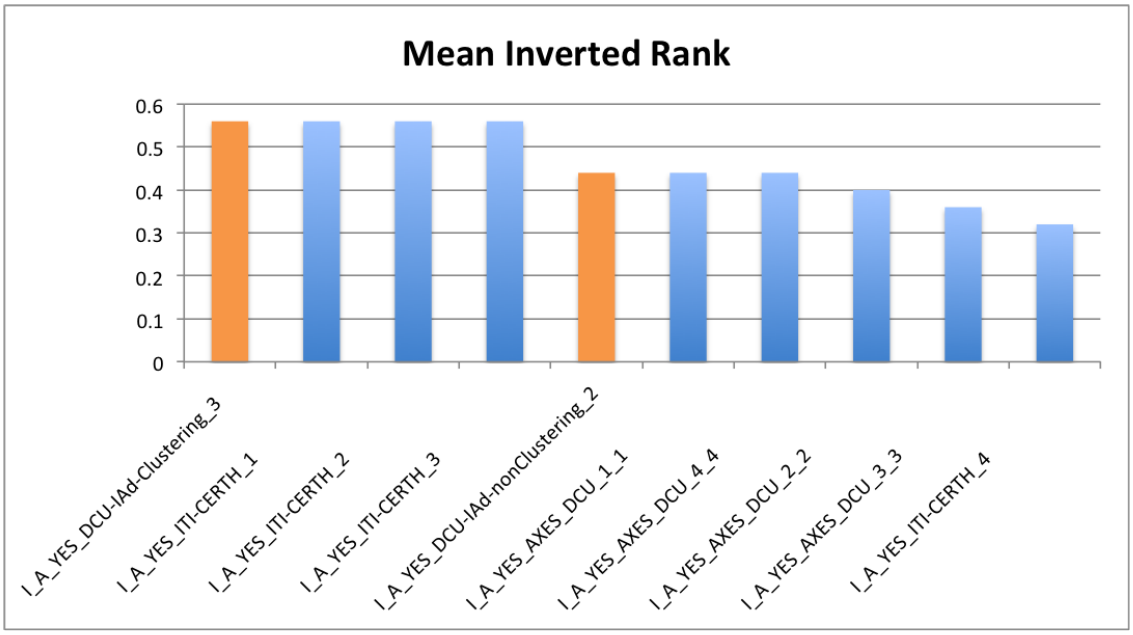


Figure 5.4: TRECVid Mean Inverted Rank results for the 10 participant runs, our submission highlighted in orange

Through our experiments, we tested whether a grouping of content based on a clustering algorithm would improve system performance over a system with no grouping, specifically for novice users. We ran these tests on the eight users which we had for the experiment, with two participants side by side running opposing systems in a competition style environment. It was clear and evident through these experiments that the clustering system was better due to participants finishing quicker. With regard to Mean Elapsed Time (see Figure 5.3), the best run with the clustering system was 2.66 minutes per topic when compared to the non-clustering system with a best run time of 3.324 minutes per topic. It was also noted that, on the clustering system, 14 topics were found where only 12 were found on the baseline system, leading to the clustering system attaining a higher Mean Inverted Rank, see Figure 5.4.

We see from the results that our clustering system attained the best results when compared to other members of the TRECVid community with respect to this Mean Elapsed Time metric. We also see that we equaled the best score with respect to Mean Inverted Rank. As far as user satisfaction is concerned we achieved an average of five out of seven for our non-clustering system and a score of six out of seven for our clustering system. User satisfaction again fell in line with topic difficulty. Users rated topics which were difficult to find requiring multiple searches with low satisfaction scores whereas those which were deemed easier to find attained the highest satisfaction rating.

As further confirmation to the success of the system, we ran experiments with an expert group of four. We found again that experts and novices performed quite similarly. These users, as was the case previously in Chapter 4, preferred to formulate their own visual queries from first search, thus bypassing the visually selected keyframe benefits. They did however find the clustering of keyframes quite useful.

5.6 Discussion

In this section we begin by discussing the clustering and non-clustering systems, examining the users search strategy. Next we look to the user feedback to gain insight into improvements which can be issued in future development. Finally we draw our conclusions for this chapter.

5.6.1 System Discussion

From the official TRECVid results above, we see that the clustering system far outperforms the non-clustering system with respect to the evaluation metrics provided through the framework. In Figure 5.5, we do a comparison of our side-by-side user tests, evaluating on each topic the number of searches performed.

In all four sets of testing, we can see that the users of our clustering system attained the best results. In the case of user 1 and 2 we see that the non-clustering system is only marginally better in two topics, with the clustering system usually

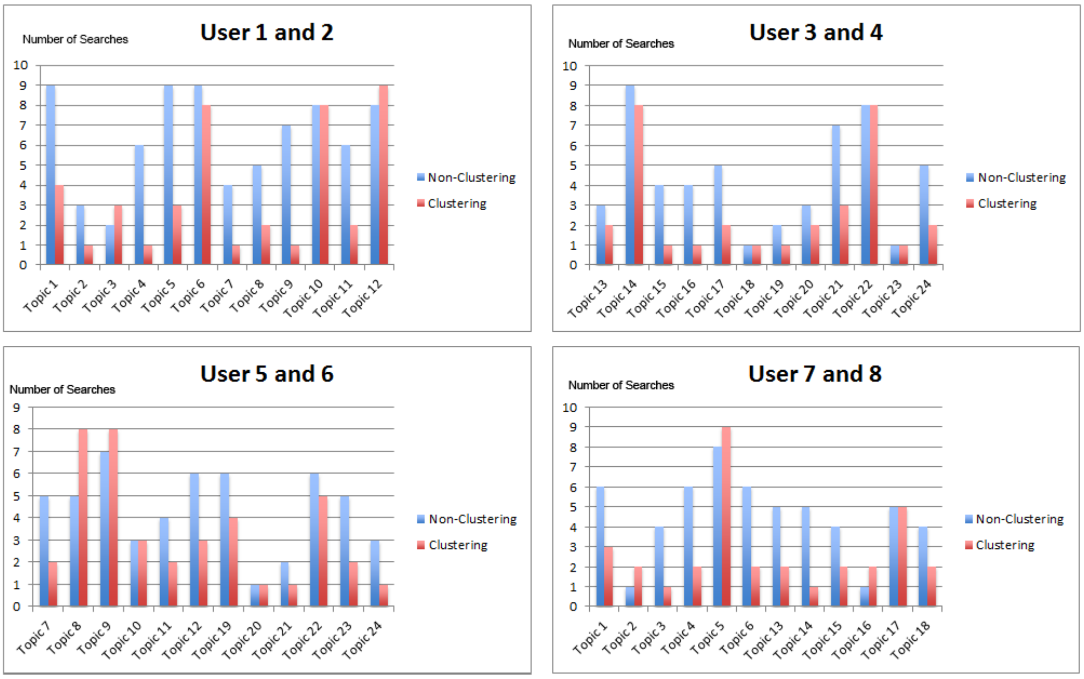


Figure 5.5: Side-by-side comparison of users based on the non-clustering and clustering systems

finding the known-item in a single search. For user 3 and 4, the baseline non-clustering system never attains better results but instead equals the clustering system three times. The clustering system is marginally better in a further two topics and much better in the remainder with most topics being found in less than three searches on this system. Users 5 and 6, showed the best results for the non-clustering this, however, was not good enough to better the clustering system showing better results in more than 50% of cases. Finally users 7 and 8, we see the non-clustering system is better in two topics, similar in a further two but performs much worse in the remainder of the topics. The clustering system in over 50% of cases finds topics in less than three searches compared to less than 20% for the non-clustering system.

In Figure 5.6 we visualize graphically the average results over the twenty four assigned topics for our eight users with respect to the number of searches carried

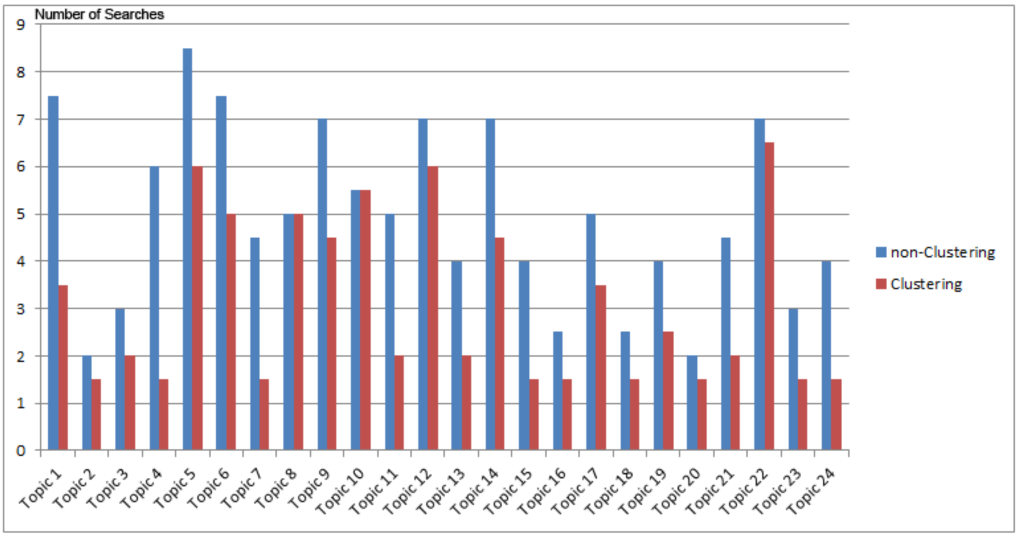


Figure 5.6: Average number of searches performed on the twelve assigned topics by our eight users

out. We see from this graph that in the majority of cases (all but two) the clustering system far out performs its non-clustering counterpart. This is further backed up in Table 5.6.1 where we see the total average searches performed of 4.9 and 3.1 searches performed on our non-clustering and clustering system respectively. Taking the 96 User Topic pairs, we find the difference to be significant based on a single tailed approach where p <0.05 defines a significant difference.

5.6.2 User Feedback

User feedback was gathered both explicitly and implicitly, each participant was provided with a paper form to fill out as part of the experimental process (see Appendix A). This feedback not only captured demographic data but also users' familiarity towards the experiment and topics. As another source of user feedback, we used the system interaction logs (see Appendix C) to determine if one system outperformed another, discussed in previous section.

	non-clustering	**clustering**
Topic 1	7.5	3.5
Topic 2	2	1.5
Topic 3	3	2
Topic 4	6	1.5
Topic 5	8.5	6
Topic 6	7.5	5
Topic 7	4.5	1.5
Topic 8	5	5
Topic 9	7	4.5
Topic 10	5.5	5.5
Topic 11	5	2
Topic 12	7	6
Topic 13	4	2
Topic 14	7	4.5
Topic 15	4	1.5
Topic 16	2.5	1.5
Topic 17	5	3.5
Topic 18	2.5	1.5
Topic 19	4	2.5
Topic 20	2	1.5
Topic 21	4.5	2
Topic 22	7	6.5
Topic 23	3	1.5
Topic 24	4	1.5
Standard Deviation	1.92	1.79
T-test		5.80E-08
Average	**4.93**	**3.08**

Table 5.12: Average number of searches per system, clustering system performs significantly better based on single tailed testing where p <0.05

The explicit user feedback was mainly positive (see Figure 5.7). Users of the non-clustering system gave a satisfaction score of the experience as a five out of seven, those using the clustering system gave it one point more at six. Users found both systems easy to use and navigate the results. With the clustering system, users liked having similar content together requiring no scrolling / extra searching to find similar content. Users found the new number of classifiers more manageable and we had more uptake in the use of concepts for searching.

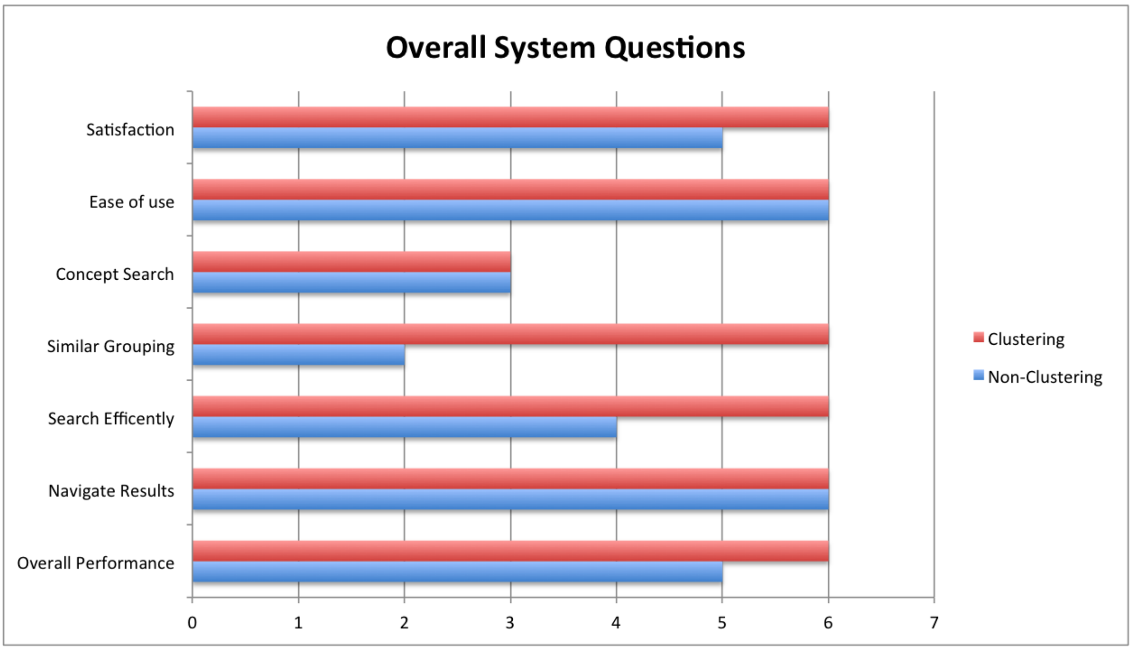


Figure 5.7: Overall system questionnaire results

However, users were uncertain as to how well the concepts performed and were confused as to how beneficial they were for searches carried out during this experiment.

5.6.3 Research Questions

To aid in evaluating our second hypothesis we devised a number of research questions

- How do we best display to the user an accurate video representation? Is a single keyframe sufficient? Can the automatic use of classifiers help with this representation?

From section 5.4 we devised multiple video representations with which we believed would aid the end user. It was found under these test conditions, that with a single keyframe representation a visually calculated keyframe based on classifier hierarchy while the preferred method, did not significantly improve on a standard middle keyframe sampling approach. For this task, users found the single keyframe sufficient to identify the video. However, we believe that adopting a multiple keyframe approach may further improve results.

- By grouping content can we provide users with a better search experience? Does a ranked list of clusters perform better than a standard ranked list?

As we can see from the previous section (5.6.2), users' prefered the clustering system over the non-clustering system in multiple different aspects. Most notably in search efficiency and grouping of content. In no case did the non-clustering system perform better, from a users perspective, than the clustering system. These results are further backed up when we show through the system logs, Section 5.6.1, that the cluster list performs better than the ranked list in terms of number of searches performed. Again this is the case in Section 5.5.1 where there is an evident improvement from one system to the other in terms of the TRECVid metrics of Mean Elapsed Time and Mean Inverted Rank.

5.7 Conclusion

In this chapter we proposed a system which has been designed with the novice user in mind, which focused on validating our second hypothesis:

> *Taking a single keyframe representation approach, where the keyframe is identified*
> *by content-based techniques, grouping similar*
> *items will help a user to more quickly locate/dismiss relevant videos.*

We have shown that, by the use of a clustering algorithm, we have reduced the amount of searches required to achieve a known item and the amount of time spent on visually processing the results from the ranked lists. We have also shown that, for this experiment, users can locate more known items with this type of system over one based upon conventional non-clustering ranked lists. Furthermore, within the scope of TRECVid, we achieved the highest rank with respect to mean elapsed time, the speed at which our system finds results, and equaled the highest number of known-items found from the community of participating Universities. Our users have high regard for this system, finding it both easy to use and rewarding in attaining appropriate search results.

The system used a single keyframe representation to help the users identify a valid video for a set of queries. One major issue with the use of a single keyframe is the system is essentially throwing away a vast majority of the semantic meaning of the video document, limiting the user to identify based on a static snapshot of data which the system assumes is relevant to the seed query. Our intention is to extend this system to incorporate greater content representation which we will outline in the next chapter.

Chapter 6

Visual Representation Comparison

6.1 Introduction

So far we have presented two experiment chapters which have dealt with developing content aware systems for the TRECVid video benchmarking conference. We focused heavily on developing for the less experience user, the novice user, who represent the demographic which access the internet with increased frequency. We believe that these users would benefit from more content based search systems. From the previous year's participation in TRECVid, we have received both positive feedback from the users and good scores with respect to evaluation metrics from the framework. For this chapter, we extend the representation of keyframes which we began in Chapter 5. While we topped the results for the known-item search task of TRECVid 2011, we believed that by ignoring the dynamic nature of the videos that we were restricting the users' view to the content. It is our belief that to accurately represent the dynamic nature of video we must be flexible in the number of keyframes displayed relating to the video. These keyframes must be semantically grouped with similar keyframes, to link like content and limit the users' likelihood of scrolling to find similar results. With this in mind, we have devised our third and final hypothesis which we will evaluate in this chapter:

Taking a multiple keyframe representation approach, we hypothesize that representing videos in a number of groups will allow for a greater opportunity in finding known-items.

To aid in the evaluation of the above hypothesis, we have devised the following research questions:

- Should we limit the number of items per cluster group? Should we merge small cluster groups which are visually close? How can we accurately determine this?
- How can we optimally represent each video and each cluster on the screen of the mobile device while showing clear distinction between cluster groups?

We began by evaluating different keyframe representation methodologies. One approach utilized extracted features to determine dissimilarity in the video. We created a representation which only contained unique keyframes. Each of the determined methods will be evaluated through user experimentation, the selected keyframe representation for each category are applied to the TRECVid 2012 dataset, iacc.1.c.

For our TRECVid 2012 experiments we reuse the evaluation framework and prototype system outlined in Chapter 3. This year we evaluate two systems; one based on the previous year′s single keyframe representation and the other based on the method of representation attained through both configuration experiments outlined in this chapter. For the first time, we do not use classifiers to aid in our official searches, but we construct a post TRECVid system which incorporates them. As ever, we evaluate our results based on both Mean Elapsed Time and Mean Inverted Rank, also examining number of searches performed per system and graphically representing these to visualize the merits of each approach.

Finally, we rerun our TRECVid 2012 experiments using a similar setup to previous years as a mark of completeness. We utilized concepts to see how we would have placed compared to other groups and to find how relevant these concepts are based on comparison with the previously run experimentation.

The chapter is presented as follows. We begin by detailing the data in the next section, paying particular attention to the differences to previous years. Next, we give an overview of the two user groups which we recruit to run both of our interactive experiments. After this, we employ configuration experiments to find appropriate keyframe representations for content of varying shot size. Then we discuss our two user testing methods, the first testing keyframe representation preferences and the second our pre-configured TRECVid system with which we field our final results for TRECVid 2012. Finally, we discuss the results and draw our conclusions.

6.2 Data

The experiments outlined in this chapter utilized the third and last iteration of the Internet Archives Creative Commons (iacc.1.c) dataset from the 2012 TRECVid data corpus. Similarly to the previous experiment, the data exhibits the same characteristics as that which has been outlined in Chapter 4. Along with the usual video documents, we are provided with meta-data relating to each video and a time-coded text transcript of the spoken word attained though the use of machine learning techniques (outlined in Chapter 3). Again we have access to the visual ground-truth which provides annotation as to the concept contained in certain test keyframes. This allowed us to develop our classification models.

The dataset is further disjointed into two separate parts, training and test. Both parts contain a mix of both professional and user generated video content. These video documents are of lengths up to 3.5 minutes with over 200 hours of test and training video alike. The training data this year has been created

Structure of the Provided Data TRECVid 2012

Hours Of Video in Test Collection	200
Number of Videos in the Training Collection	8216
Number of Shots in the Training Collection	137327
Number of Videos in the Test Collection	8263
Number of Shots in the Training Collection	145634
Number of Topics in the Training Collection	300
Number of Topics in the Test Collection	24

Table 6.1: Overview of the data used for TRECVid 2012

from the previous year’s test data, which has undergone the usual post TRECVid experiment analysis (see Chapter 5). The data is represented in Table 6.1.

6.2.1 Additional Data Sources

Three additional sources of data which aid us in developing visual search components with which we used to develop our system were:

- Classifier Ranked List
- Multi Keyframe Video Representation List
- Multi Keyframe Cluster List

We created classifiers again for the seven most commonly used concepts, as discovered from chapter 4, using the extracted feature attained though OpponentSIFT (a color biased feature). OpponentSIFT was chosen due to its good performance which had been outlined by van de Sande et al. (2011) and van de Sande et al. (2010). We again use an SVM to train the classifiers based on positive and negative examples attained through the visual ground-truth. A ranked list is generated for each of the trained classifiers based on the probability of containing the concept. These lists form the basis of the classification search in the final system.

We use the MPEG-7 features of Edge, Scalable color and color Histogram to represent a calculated multiple keyframe representation. By calculating the image similarity of the keyframes local to each video we can remove near duplicates and provide a set of unique representative keyframes for each video document in the collection.

Finally we employed the visual features in conjunction with a clustering algorithm. In this way, we provided a ranked list of clustered documents. Each cluster can contain information from a video document once, but video documents can belong to multiple clusters. As such, the same video can be represented based on any of its keyframes and increased in rank based on the clustering approach. We will discuss this in more detail in Section 6.5.

6.3 Users

In order to evaluate the our experimental designs, we ran three sets of user experiments. The first experiment gauged user preference towards multiple representations of different video documents. The next experiment evaluated our configured system for TRECVid 2012 known-item search task. Our final experiment reran the experiments of TRECVid 2012 with a different user group to compare the initial run which used no classifiers with a new run that did. As such, we will required three distinct groups of novice user to test each of the criteria, outlined in Table 6.2.

- **Experiment 1:** We recruited six participants to run this experiment. These students again featured a majority of undergraduate students who had recently completed third year In a university level degree program. All participants were under 25 years old and were heavy web user with respect to both video and text search, they all came from an undergraduate program in either a science or engineering background. None of the members had ever

Participant Profile		Exp 1	Exp 2	Exp 3
Age:	18 -25	6	0	0
	25 - 30	0	8	4
Web Search (inc Video)	Regular	6	7	4
	Infrequent	0	1	0
Handheld Usage	Never	0	1	0
	Infrequent	1	5	2
	Regular	5	2	2
Education	Undergrad	6	1	4
	Graduate	0	7	0
Gender	Male	4	3	0
	Female	2	5	4

Table 6.2: Participants Profile for User Experiments

been involved in the creation of a system similar developed for TRECVid or otherwise and are ideal representatives of a novice user group.

- **Experiment 2:** Due to limitations with the wind-down in the project we were unable to recruit participants from our Norwegian partners, instead we look to members of our wider research group that fit the profile of novice users, namely those who have not been affiliated with creation of search systems similar to the TRECVid model. We recruited eight member to test for this experiment. All members were over the age of 25 with the majority being graduates. Most of the users would regard themselves as medium internet users, using Google and YouTube quite frequently.

- **Experiment 3:** Post TRECVid experimentation revealed a gap in our research model, we recruited four users to rerun a single iteration of the experiments on a modified system. The recruits fell into the category of novice users. All four were final year nursing students who engaged in light to medium internet searches. All participants were over the age of 25. These participants engaged in regular light usage of Google, Facebook and YouTube. Outlined in Table 6.2.

6.4 System Configuration Phase

Before deploying our final system for TRECVid 2012, we first devised two small scale experiments to aid in configuring the system with respect to multiple keyframe representations. We began by identifying categories with which our content could be logically placed, for example videos with less than five keyframes up to those with more than fifty. We developed a representation based on the extension of the single keyframe representation seen in Chapter 5, using this frame as a stem to calculate dissimilar keyframes for representation recursively. This visual calculation was applied to each video and subjected to a round of user testing. In the following sections, we discuss creation of the visually calculated keyframe representation and the user testing to indicate preference of keyframe representations.

6.4.1 Multiple Keyframe Visually Calculated Representation

Our first task was to determine an automated approach for multiple keyframe representation for videos in the TRECVid corpus. As the feedback from the pilot task in Chapter 4 noted keyframe duplicates were unacceptable, we used a single keyframe representation in the last chapter to alleviate this problem. This single keyframe representation restricted the dynamic nature of the video document by limiting the amount of viewable content, thus reducing the likelihood of finding the known item. To combat this lack of detail in the representation, we once again looked to multiple keyframe representations. To build this multi keyframe representation we required a single keyframe to seed from. We achieved this in Chapter 5 by finding an appropriate single keyframe representation. We decided to use the central keyframe as opposed to the visually calculated keyframe due to the lack of significant difference between the two and the extra processing overhead associated with the visually calculated frame. Having a blanket keyframe

Category	< 5	5 - 10	11 - 20	21 - 50	> Fifty
Amount	3572(43.2%)	1299(15.7%)	1414(17.1%)	1422(17.2%)	556(6.7%)

Table 6.3: Video shot category groups and number of videos of the 8263 which fall into each category

representation for all videos makes little sense as the videos can fall into different categories. We identified five such categories with which the videos could be grouped (see Table 6.3). Each category had the potential to have a different representation associated with it.

Selecting twenty five exemplar videos, five from each of the categories, we extracted MPEG-7 visual features (edge, color histogram and scaleable color) for each shot. Using the distance formula we found the furthest keyframe vector representation from the initial seed frame. If the keyframe was outside of a certain threshold (dissimilar), we add it to the set of representative keyframes. We continue by calculating the distance from each frame in the set to next furthest keyframe. If this next furthest frame is past the threshold for each of the contributing frames, we again add to the set and repeat until no more frames satisfy the criteria. Once achieved we have our final representation for that video.

Results

From the experiment outlined above we gain information on an effective representation based on videos of different shot sizes (see Table 6.4) . We found that videos with over fifty shots, a small percentage of the data, usually represent videos containing image bursts. These large representation videos exhibit high dissimilarity in the videos content. We see this by the average number of representative keyframes being 111.6. From the set of videos with less than five keyframes (almost 40% of the collection, of which 1421 or 17% are single shot videos) we saw an average of 1.2 frames.

Category		< 5	5 - 10	11 - 20	21 - 50	> Fifty
Video 1	Shots	1	7	16	25	200
	Relevant	1	2	7	10	123
Video 2	Shots	3	6	19	30	151
	Relevant	2	3	5	7	55
Video 3	Shots	1	8	12	43	739
	Relevant	1	3	4	9	315
Video 4	Shots	4	6	13	23	210
	Relevant	1	2	4	9	48
Video 5	Shots	2	9	17	29	75
	Relevant	1	4	5	11	17
Average		1.2	2.8	5	9.2	111.6

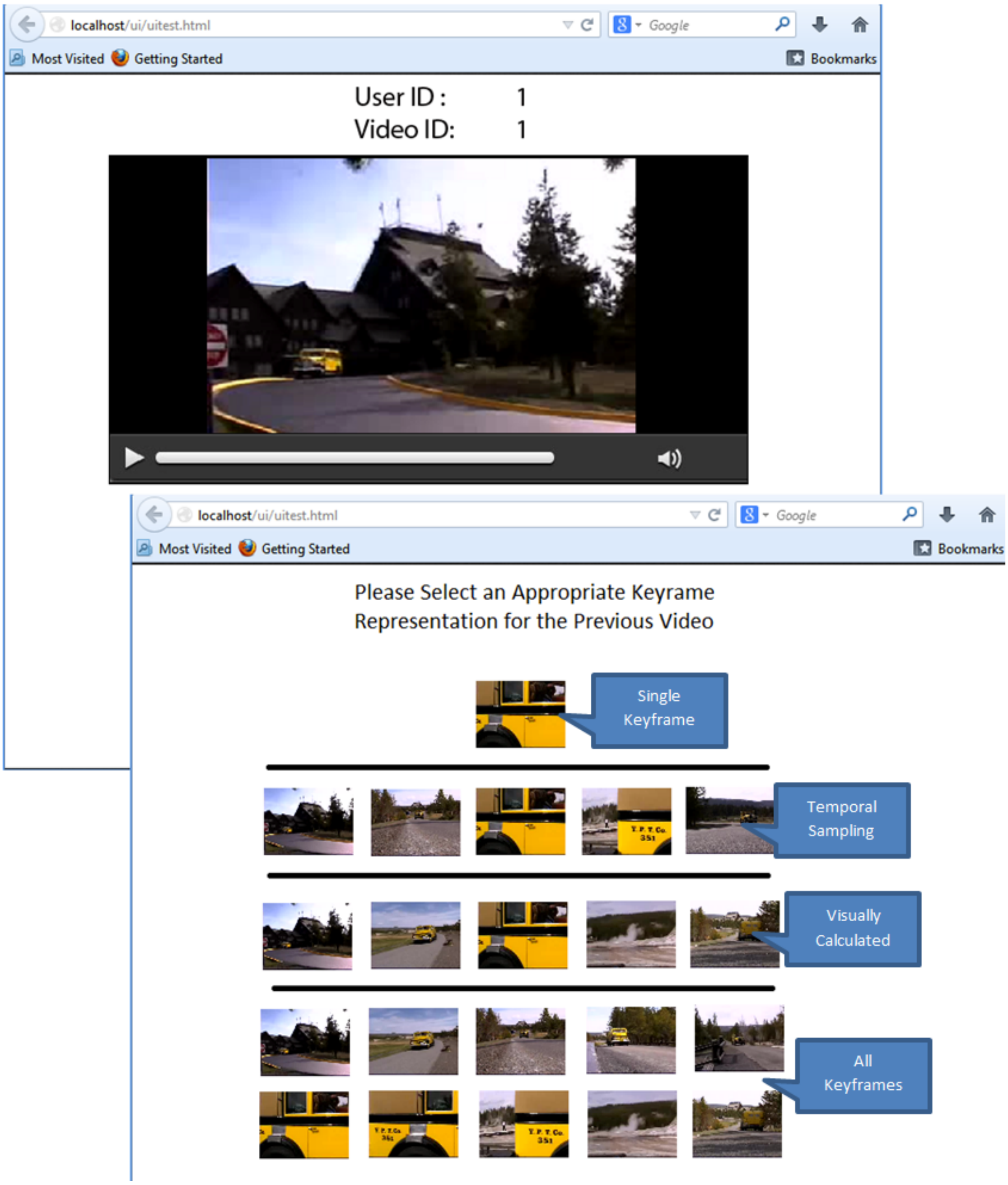
Table 6.4: Result table from automatic experiment to detect dissimilar keyframes from within videos

The content in the other groups featured both user generated and professional content such as news reports, game shows and interviews. As such it had multiple shot representations, some of which featured duplicate shots. With an average keyframe representation over the three remaining categories being approximately five frames, it is in this central grouping where we believe our approach will have the most effect.

The next section is concerned with the user experimentation carried out in configuration of the 2012 TRECVid system. We begin by running small scale test to determine user preference towards different types of representation on each of the categories defined earlier. One of the methods tested is described in this section.

6.4.2 User Evaluation of Multi Keyframe Representation

In the previous section, we utilized visual features to determine a keyframe representation for videos of varying shot sizes. In this section, we evaluate through user testing this representation compared to alternative approaches. Videos which contained less than five shots were not evaluated by user testing



User ID : 1
Video ID: 1

Please Select an Appropriate Keyrame Representation for the Previous Video

Figure 6.1: Experimental interface for keyframe representation experiment

as we deemed a single keyframe representation to be sufficient, as almost half of the videos in this group contained only a single shot. We have determined four methods with which to present to our users for testing outlined in the list below.

- Single keyframe representation, for videos containing more than five keyframes, a baseline approach selecting the central keyframe to represent the video, as seen in Chapter 5.
- Temporal sampling, excluding the extreme cases of both the five or less and fifty and more shots. We found the average number of representative keyframes for the remaining categories to be 5 frames. This method attempted to mimic the visually calculated approach with respect to returned frames but with a much decreased processing time. To populate this representation we used the central frame, first frame and last frame. The remaining two frames were attained from the halfway point of the central to the first frame and from the central to the last frame.
- Visually calculated representation, based on the method outlined in the above section.
- All keyframe representation, returns all keyframes relating to video. This method we believe will cause multiple duplicates.

For this experiment we reused the system developed for attaining user preference as seen in Chapter 5. Each of the representations above were integrated into the system (see Figure 6.1). We recruited six novice users to run the experiments. Participant profiles are explored in the user section above. We began by giving a brief demonstration of the system informing the users they would see a short video. Once the video had completed they would be presented with four options with regard to representation and they could pick their preference. Users were then asked if they would like to begin the test or take a training topic. Most chose to take the test as they believed the demonstration to be sufficient. We used 24 videos from the training set, choosing different videos from those in the previous section. Each user was assigned 12 videos, 3 from each category, and given 60 minutes for the entire task, allowing 5 minutes to watch each video and make a decision with respect to preference. Each video had coverage by three users which can be seen in the results Table 6.4.2.

Results

The results from the user testing are outlined in Table 6.5. We can see that our visually calculated keyframe representation method is by far the most popular with 41.2%. The next most popular was the Temporal sampling representation with 36.1% and finally with the single keyframe representation, which presented so well last year, only being preferred in 17% of cases. It was noted that the fourth method of representation, that of using all keyframes, was rarely selected. Users stated this was due to multiple keyframe duplication.

6.5 TRECVid 2012 Experiments

6.5.1 TRECVid 2012: Single vs Multiple

The first experiment was carried out utilizing the final dataset (iacc.1.c) from the known-item search task framework of TRECVid 2012. We rebuilt the system using our prototype which was outlined in Chapter 3 and the new data described earlier (see Figure 6.2). Eight novice users attained from adjacent research groups formed the participants for our experiments. These users were assigned twelve tasks each, 6 topics on single keyframe and 6 on multiple keyframe, the distribution of which is shown in Table 6.5.1. They were also provided with instructions on how to use the system and a set of training topics which were run to familiarize them with

	User 1	**User 2**	**User 3**	**User 4**	**User 5**	**User 6**	**Shots**
Video 1:	3		3		1		5 - 10
Video 2:	3		2		2		11 - 20
Video 3:	3		2		3		21 - 50
Video 4:	3		2		3		>50
Video 5:	1		1		1		5 - 10
Video 6:	2		2		2		11 - 20
Video 7:	2			3		3	21 - 50
Video 8:	3			2		3	>50
Video 9:	1			3		1	5 - 10
Video 10:	3			2		3	11 - 20
Video 11:	3			2		3	21 - 50
Video 12:	3			4		2	>50
Video 13:		2		3	2		5 - 10
Video 14:		3		3	2		11 - 20
Video 15:		3		3	3		21 - 50
Video 16:		3		2	2		>50
Video 17:		1		1	1		5 - 10
Video 18:		2		2	2		11 - 20
Video 19:		2	2			3	21 - 50
Video 20:		2	3			3	>50
Video 21:		1	1			2	5 - 10
Video 22:		2	2			3	11 - 20
Video 23:		2	3			1	21 - 50
Video 24:		4	4			3	>50

Table 6.5: Table outlining user preference of keyframe representation, 1.) Single Frame 2.) Average Number of Frames 3.) Visually Calculated Frames 4.) All Frames

the system. Finally, users were given a survey form which they completed at each stage of the experiment; pre-experiment to capture demographic and usage data, post-experiment to capture their overall perception of the test and a survey at the end of each of the 6 assigned topics to see how each representation fared.

We again employed a clustering technique to group similar content, k-means with *k*, the amount of clusters, set to 100. This value was defined from experiments outlined in Chapter 5. Both systems use the output provided by this visual clustering, as such the only element we test is the single vs multiple keyframe representation. We were unable to represent multiple keyframes in the traditional way (storyboarding), where videos are represented together on the same line. In this system the multiple keyframes each belong to a separate clusters, allowing for increased likelihood in finding relevant videos quicker than with entry in a single group, as was the case in TRECVid 2011.

Results

In this set of experiments, we omitted the use of visual classification, in part to test the effectiveness against against other members in the community. This, however, proved costly as, due to this lack of classifiers, both automatic and explicit we only found ten of the twenty four topics on the single keyframe system and a further two were found in the multiple keyframe system. This is a stark contrast from last year's experiments where we attained the best results finding fourteen of the known-items on a single keyframe representation system which utilized classification.

With regard to Mean Elapsed Time, our multiple keyframe approach out performs the single keyframe representation by almost a minute (see Figure 6.3). In terms of Mean Inverted Rank, we also witness that users of the multiple keyframe system perform better finding more known-items. Overall our results rank in the middle for the multi-keyframe representation to the bottom for the

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

Table 6.6: Table outlining the topic distribution over our eight participating users, (a) denotes the Single-Keyframe system (b) denotes the Multi-Keyframe system

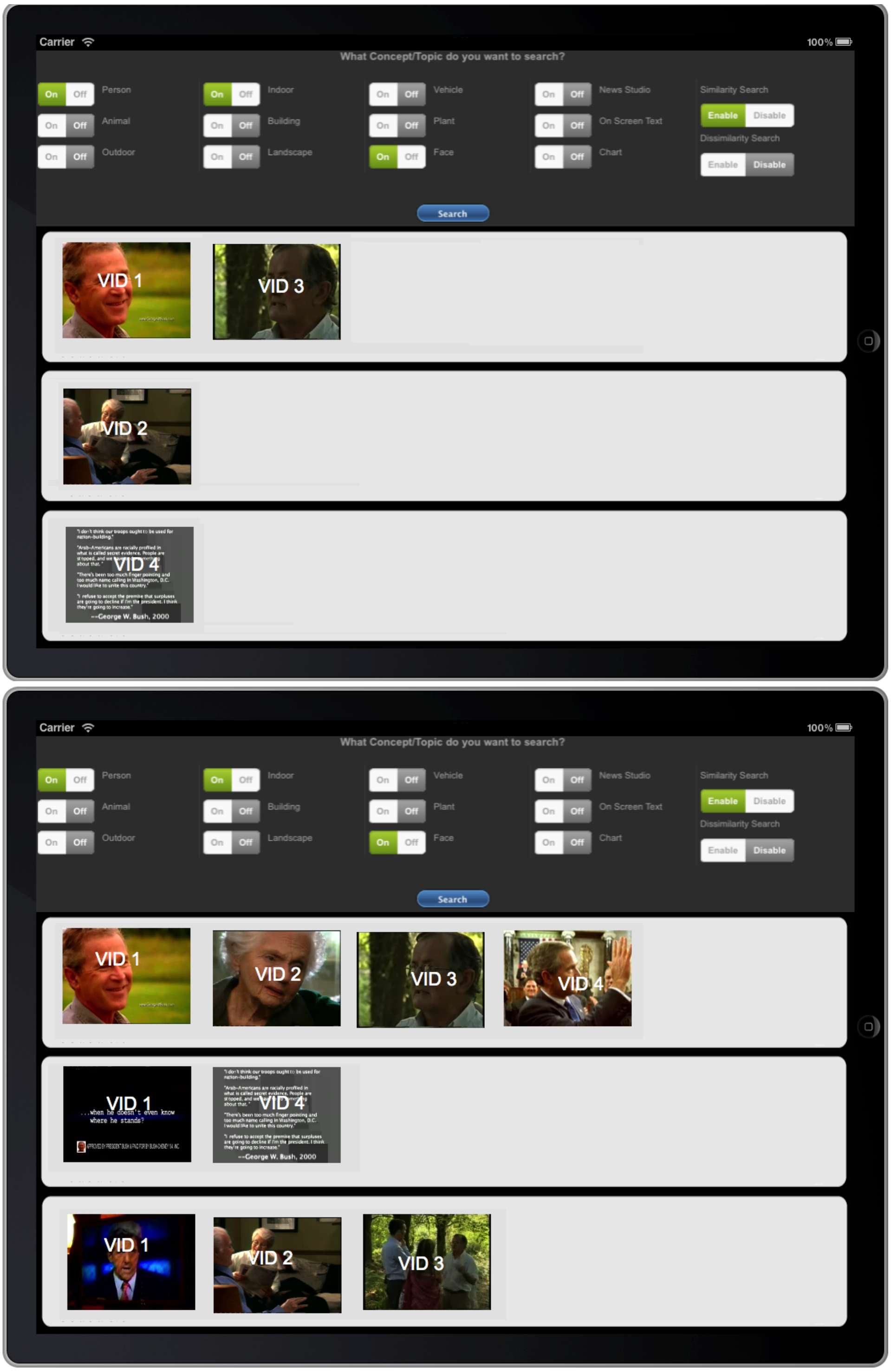


Figure 6.2: A look to the interface used in the 2012 TRECVid experiments. We used a HTML5 based interface but kept the interaction the same as previous years. This screenshot shows the top single clustering system with only one keyframe per video hit and the multiple clustering system with different keyframes from the same video belonging to different clusters

]

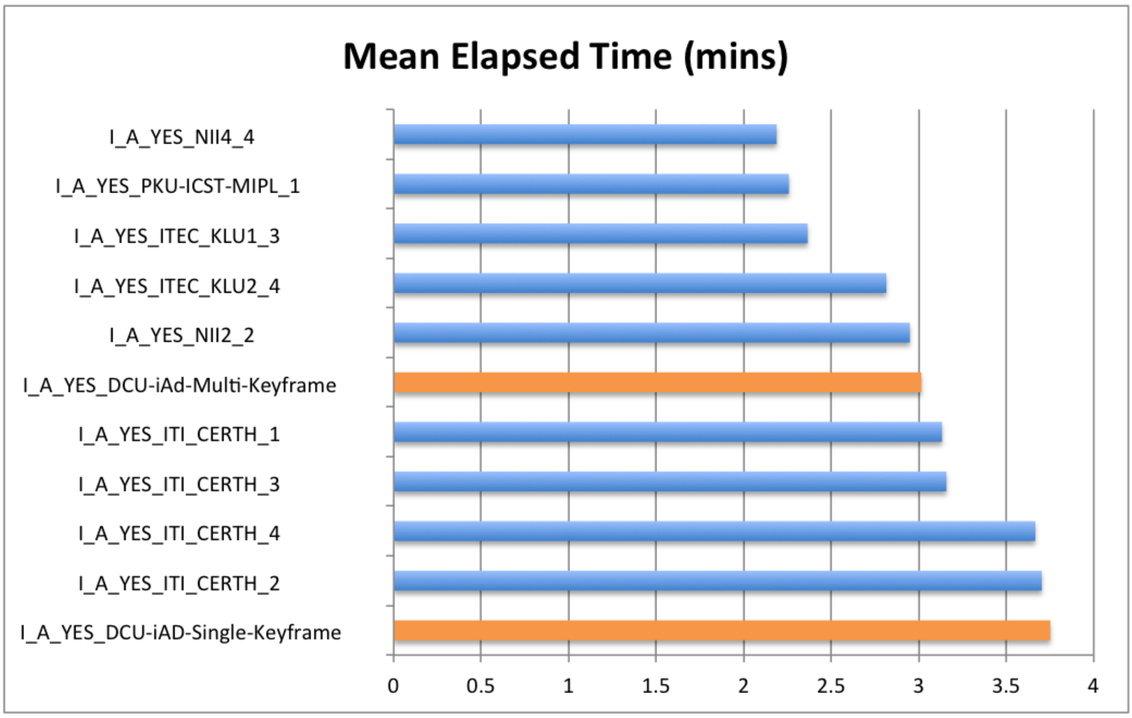


Figure 6.3: Mean Elapsed Time of teams participating in TRECVid Known-Item Task, our submission highlighted in orange

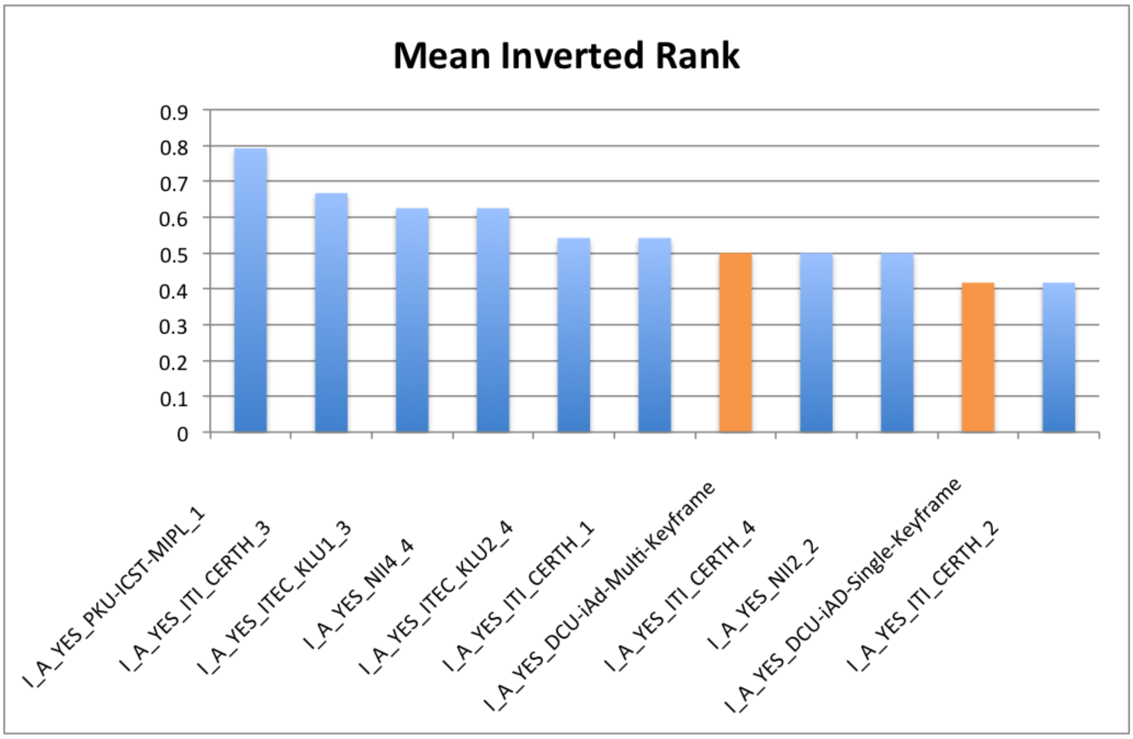


Figure 6.4: Mean Inverted Rank of teams participating in TRECVid Known-Item Task, our submission highlighted in orange

single keyframe representation. From comparison with the groups which used classifiers, we see from this that the classifiers do in fact help greatly with allowing the user to find an effective known-item to satisfy the topic description.

6.5.2 TRECVid 2012: With Concepts

We found through post TRECVid analysis that our system under-performed when compared with members of the research community who also submitted to the known-item task. We decided to rerun a single iteration of experiments on the system utilizing seven trained classifiers which were developed over the training data from TRECVid 2011 and applied to the test data for this year. We recruited four new novice users to act as our users for this experiment, who were assigned topics as seen in Table 6.5.2. For each of the participants we record both demographic user data including familiarity towards topics and an exit survey which capture the users ranking for the experiment.

As with previous versions of the TRECVid style experimentation we gave a demonstration of the system to the users showcasing each of the systems features and how to test for known-items. We also provided users with a number of training topics to familiarize themselves with the system functionality. All of our users were given a formal instruction list which outlined the topics to be tested with a textual description of the video to be found. Users were only asked to proceed with the experiment when they were comfortable and after running a number of training topics.

Results

We evaluated this experiment based on the best overall combined result for each of the assigned topics, in this way we found 15 of the 24 known-items leading to a Mean Inverted Rank of 0.625 (see Figure 6.6). We also witnessed a Mean

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

Table 6.7: Table outlining the topic distribution over our four participating users for the multiple keyframe representation approach featuring classifiers

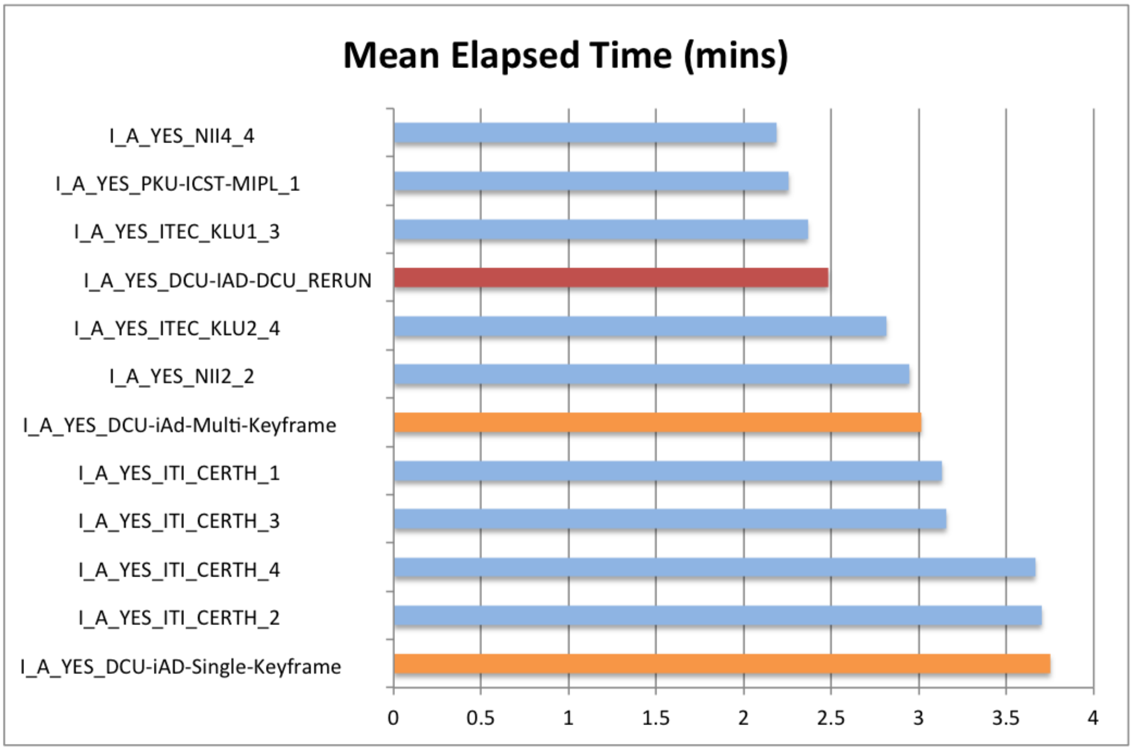


Figure 6.5: Mean Elapsed Time of teams participating in TRECVid known-Item Task, our submission highlighted in orange, rerun highlighted in red.

Elapsed Time of approximately 2.5 minutes per topic, 30 seconds quicker to find the known item than without using classifiers (see Figure 6.5). Overall, our rank increased from fourth to third place, a 16% increase in Mean Inverted Rank and a 17% reduction in Mean Elapsed Time.

6.6 Discussion

In this section we drill down further into the results attained through the three sets of user testing relating to the TRECVid 2012 experiments.

6.6.1 Result Discussion

From our official results, we see that not only does the multiple keyframe clustering approach out-perform the single keyframe clustering approach in terms

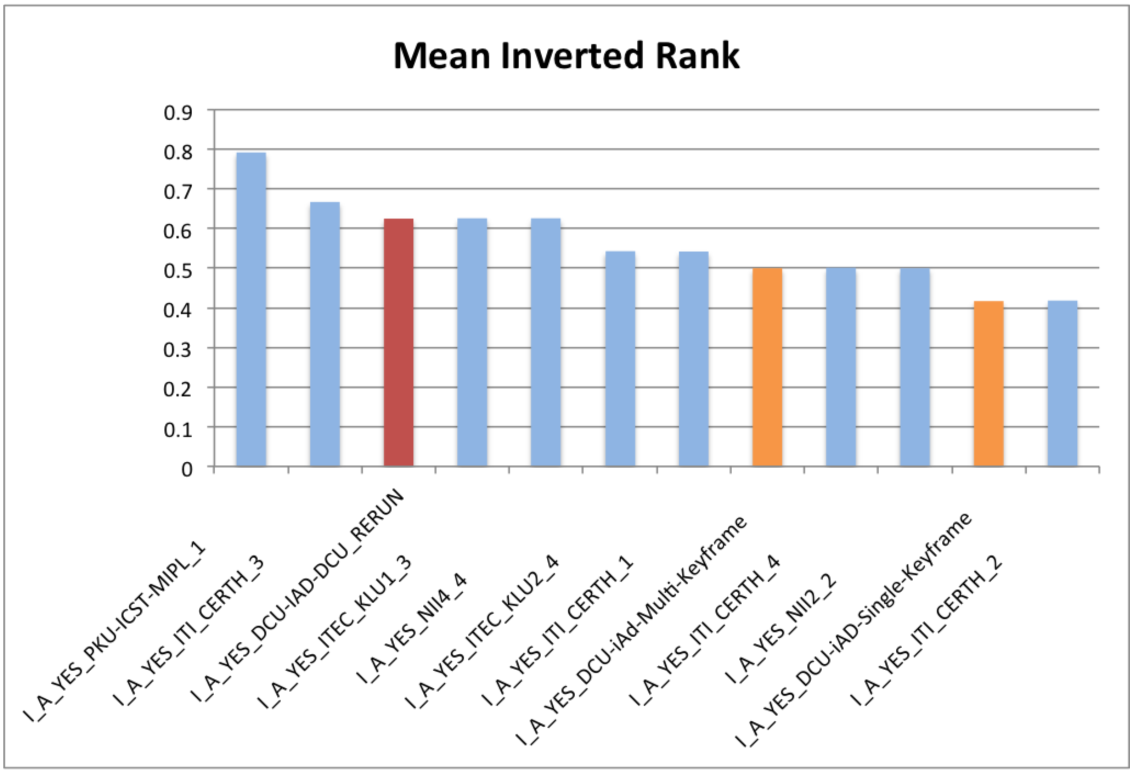


Figure 6.6: Mean Inverted Rank of teams participating in TRECVid known-Item Task, our submission highlighted in orange, rerun highlighted in red.

of time to find the item by almost a minute, but users found more of the known items in the allotted time window. Looking into the users' search strategy for both approaches (see Figure 6.7), we notice that in all cases of the multiple keyframe approach, user perform fewer searches to find the known item. In the case of user 1 and 2, we see four clear cases where the known item is found with fewer searches than the single keyframe system, with the remainder being found in a similar amount of searches. Topic 8 appears to be a random event and associated with the user's miscomprehension of the topic due to being a non-native English speaker. With users 3 and 4, again we see four clear cut examples where the multiple keyframe approach performs better and only a single case where the reverse is true. We see another outlier here in the form of topic 9. The video associated with this topic featured no metadata, it was boosted in rank by the clustering system to appear earlier in the results list. Users 5 and 6 show only

	Single	Multiple	With Classifiers
Topic 1	6.75	5.75	3.25
Topic 2	5.5	5.25	1.5
Topic 3	4	3.25	3
Topic 4	5.25	3.25	3.25
Topic 5	6.5	6.5	5.75
Topic 6	5	5	4.25
Topic 7	5	4.75	4
Topic 8	4.5	2	2
Topic 9	3.75	1.75	2.5
Topic 10	6.25	6.25	5.75
Topic 11	5.25	5	3
Topic 12	4.75	4	2.5
Average	5.21	**4.4***	**3.4***

Table 6.8: Table outlining the average search by users over the twelve assigned topics, Multiple representation shows significance over single and Multiple with classifiers shows significance over Multiple without Classifiers (single tailed p <0.05)*

three cases where the results diverge from the single keyframe representation. With users 7 and 8, we witness for a third time four instances where the multiple keyframe representation requires fewer searches to find the known-item. Averaging the number of searches over the entire collection, we are presented with Table 6.8. Here we see that the average number of searches for the single keyframe system is 5.21 with the multiple keyframe system requiring only 4.4. searches. Using a single tailed t-test where p<0.05 on the 48 <*User, Topic*> pairs revealed a significant difference between each of the proposed approaches. Since we are using paired t-test, the 48 pairs is over the minimum N of 30 with which we can define significance.

Post TRECVid experimentation revealed a topic for additional analysis. Having not tested the second approach with a similar setup to previous years we could not accurately assume that one approach was better or worse without first giving this system the advantages of using classification models. We see from this testing that not only did our user base find more of the known-items but

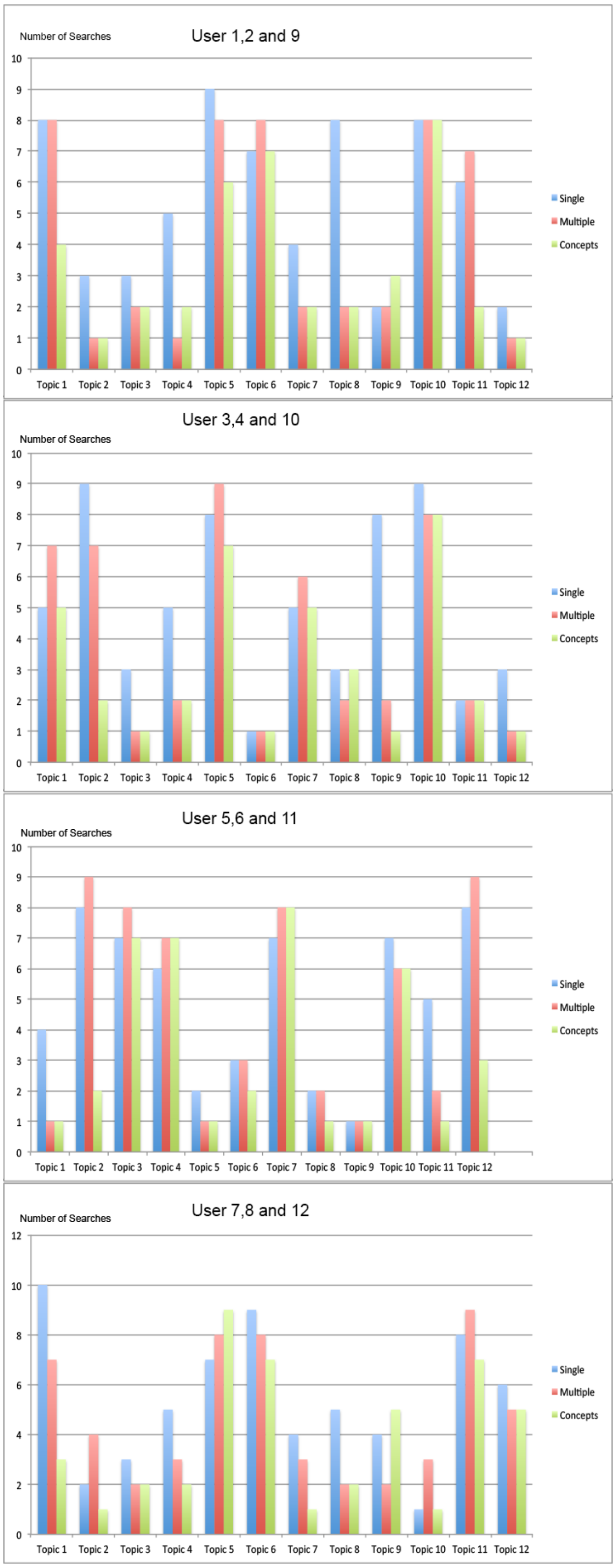


Figure 6.7: Visualization of the number of searches performed in each of the four test sets

they were quicker at finding them. In Figure 6.7 we see a visualization of the searches performed when compared with the previous experiment. In all cases with the use of classifiers the users performed as well or better than the multiple keyframe without classifiers. It was noted that in most cases, users didn't use classifiers until their second or third search. We see when taking the average searches performed that users perform on average 3.4 searches per run, in which case the items are found while a user has issued some classifiers. We also found by running a similar t-test on this and the set relating to multiple keyframes without classifiers that there was a significant difference.

6.6.2 Result Comparison

In this section we look at the three years of participation in TRECVid Known-Item search. We analyze the results with respect to Mean Inverted Rank and Mean Elapsed Time for these years. We directly compare the results each year to the average performance, looking at how the systems evolved over the years with a view to reducing the search overhead .

Mean Inverted Rank

In Figure 6.8, we see a graph depicting the scores attained through Mean Inverted Rank (MIR) by our group for the years 2010, 2011 and 2012. We also present the average rank attained through these years of all participating teams. From the figure, we can see that in terms of MIR performance there is only a slight increase in performance between the first and final year, attributed to the numerous "Hard Topics". These hard topics were related to videos which exhibited both little/no meta-data, and which ranked low with respect to visual classification. From the graph, we show that each year we are ahead of the average by 0.1 except for the

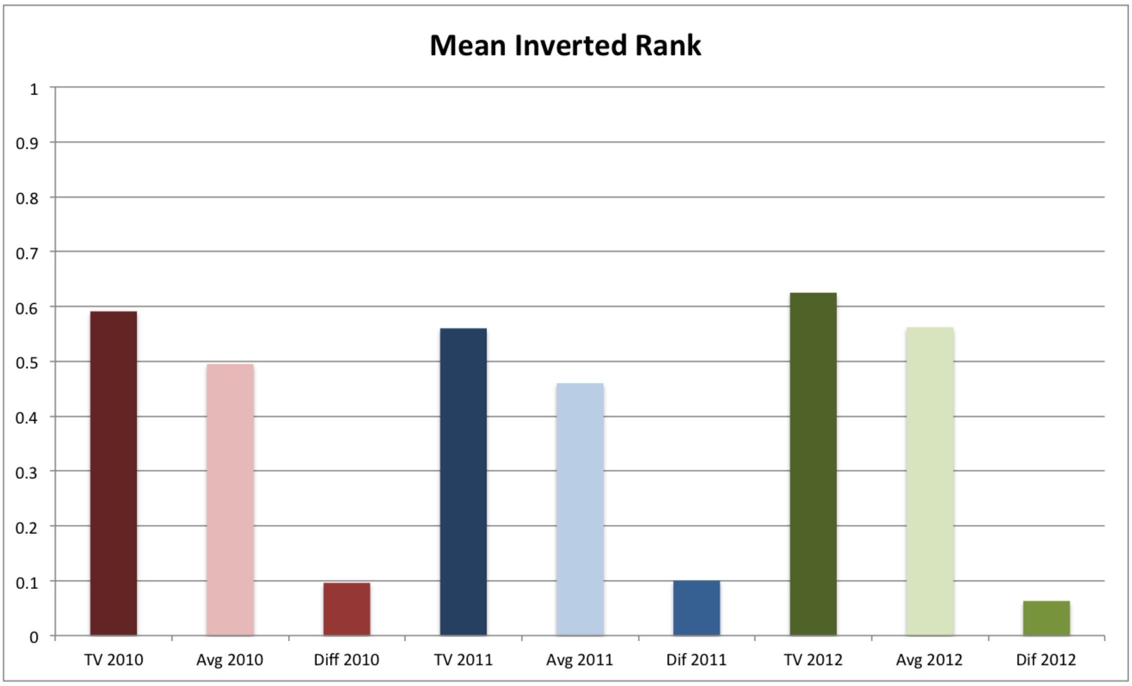


Figure 6.8: Mean Inverted Rank of our runs from TRECVid 2010 - 2012 compared to the average each year and highlighting the difference compared to the average (higher is better)

final year, we believe the gap to be closer due to other teams learning successful approaches from the teams attaining top results in the previous years.

Mean Elapsed Time

In Figure 6.9, we see a similar graph this time showing the Mean Elapsed Time with respect to the three years of participation in the known item search task. In this graph, we can see that our users' interaction time with respect to finding items has decreased each year by 0.25 minutes. We also note from this, that on average the other participating groups' interaction time gets slightly worse or stays almost the same. We see from this, in the 2011 experiments, that our system performed tasks 0.75 minutes faster than the average of all participants, a stark contrast to the previous year where we were only marginally faster.

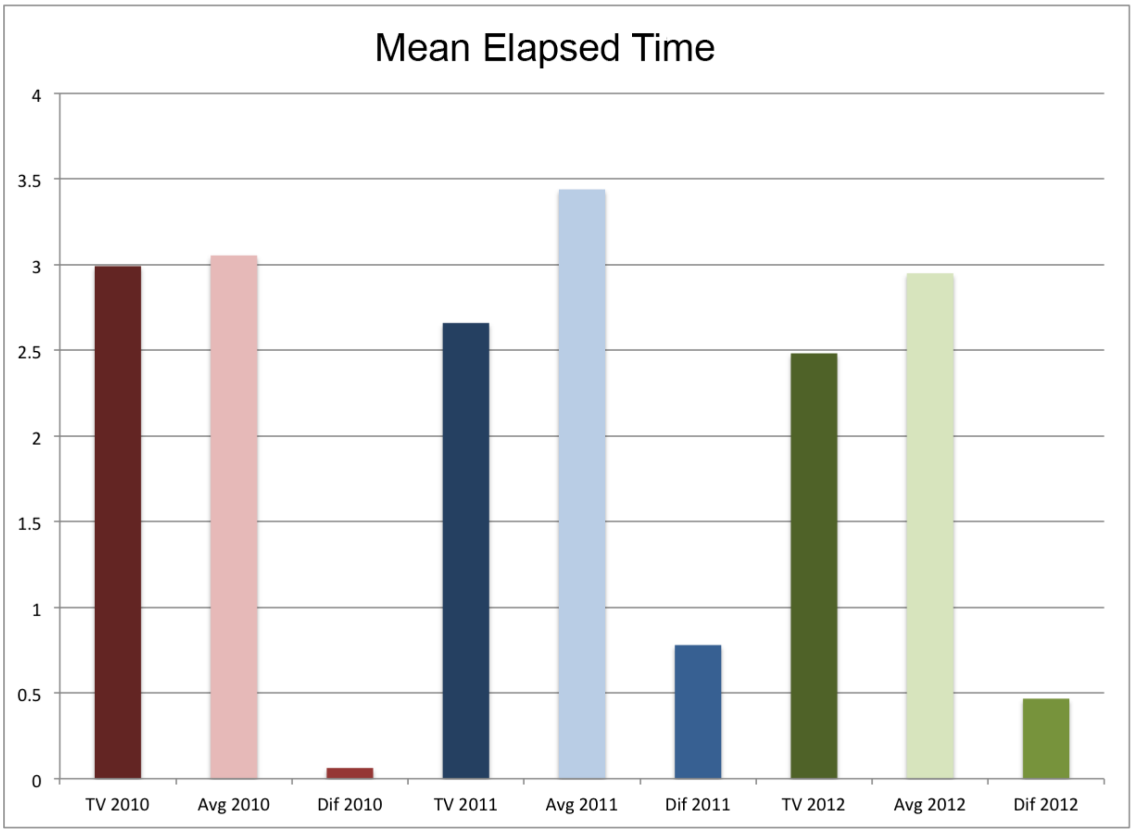


Figure 6.9: Mean Elapsed Time of our runs from TRECVid 2010 - 2012 compared to the average each year and highlighting the difference compared to the average (lower is better)

6.6.3 Research Questions

- Should we limit the number of items per cluster group? Should we merge small cluster groups which are visually close? How can we accurately determine this?

We decided not to limit the number of items which belonged to each cluster, if an item logically belonged with other similar items it was placed with them. We also chose not to merge smaller cluster groups, to give a clear distinction with in the clusters, this is an avenue which can be explored as part of future work.

- How can we optimally represent each video and each cluster on the screen of the mobile device while showing clear distinction between cluster groups?

We distinguish each cluster by limiting it to a single line on the interface. While this does cause certain areas to be associated with blank canvas, it allows users to more easily accept/dismiss content without wondering does the next line also belong to this current cluster.

6.7 Conclusion

For this our final experiment chapter, we proposed a system which utilized a multiple keyframe representation to increase the likelihood of finding items of relevance. Instead of using the traditional story-board method we deploy unique frames into different clusters. This is outlined in our third hypothesis:

Taking a multiple keyframe representation approach, we hypothesize that representing videos in a number of groups will allow for a greater opportunity in finding known-items.

Leading on from the success of 2011 we adopted the same clustering algorithm. This time, however, we applied it to a multiple keyframe representation. We believed by giving the users more opportunity to view the content in a different light, the higher the likelihood of them finding the known item.

We have seen from the above work, that in terms of both the evaluation outlined by TRECVid and the average number of searches performed, that overall our clustering system on multiple keyframes outperforms that of the single keyframe clustering variety.

We did not include concept search in our initial submission to TRECVid and this led to mid-table results. Post experimentation with the same system with concepts active yielded much higher results. This shows that the use of concepts can be quite beneficial and should be addressed in future work.

While the multiple keyframe system performed better than the baseline system of previous years, the lack of support for classifiers initially reduced the search power of the user. This clustering support, especially its usage by novices, could provide for further research in the area.

Chapter 7

Conclusion

In this concluding chapter, we summarize the presented work, paying particular attention to how it has aided in proving our hypotheses. Next, we look to the contributions and shortcomings of this work. Finally, we look to what the future may hold within the scope of this work, suggesting possible future work which could further enhance our retrieval methods.

7.1 Summary

Within the field of video retrieval we have witnessed a huge growth in the amount of content being published in the last few years. This growth is attributed to devices such as the iPad, which have allowed for a steady stream of new content continuously being uploaded to WWW archives. These larger scale content systems are beginning to require content-based methods to classify this enormous quantity of data and to give it some sort of structure. User Generated Content, a relative unknown 10 years ago, now plays a major part in the everyday life of the average web user, from the daily Facebook digest to the YouTube video recommended by a friend. We have witnessed a new trend in media production where everyone can be both a content producer and content consumer. This thesis was concerned with evaluating approaches which aided the user in querying content based systems from a non-expert perspective. In the rest of this chapter we summarize our work.

In the first chapter, we began by giving a brief description of the area of information retrieval (IR) paying particular attention to video retrieval, the chosen topic of this thesis. Beginning by discussing what motivates this research; with large archives trying to find items of relevance is getting harder, some have little/no meta-data and require content-based techniques to gain semantic meaning. This type of content discovery requires complex querying which standard (novice) users are unfamiliar with. As such, there is a gap to integrate content based techniques seamlessly into large scale search archives, giving casual users the power to search at a similar level to experienced users.

For this work we created three hypothesis which helped with addressing the problems stated in our motivations:

1. Using a tailored interface design, which utilizes selected content based retrieval techniques, on handheld devices will increase the performance of novice users when carrying out known-item search tasks.

2. Taking a single keyframe representation approach, where the keyframe is identified by content-based techniques, we hypothesize that grouping similar items will help a user to more quickly locate/dismiss relevant videos.

3. Taking a multiple keyframe representation approach we hypothesize that representing videos in a number of groups will allow for a greater opportunity in finding known-items.

Next, we looked at literature within the field of information retrieval, focusing on areas which directly relate to the work in this thesis. We analyzed the literature from a high-level, with emphasis on research carried out by groups participating in the annual TRECVid video retrieval conference. We paid particular attention to interface design, visual keyframe representation, and clustering algorithms all of which were of paramount importance for our work. In this chapter, we provided a historical overview of TRECVid, from the early days while still associated with text retrieval, to more modern times where it contains multiple tracks, all with different topics and delivery criteria. This large scale video benchmarking conference has helped push the envelope with respect to research systems. Finally we give an overview of the participation in the chosen TRECVid task of known item search and how our research fits in with the rest of the community.

In Chapter 3, we began by describing the system used to support our interactive experimentation, that of the Clipboard system. This system was developed with a modular structure, to be easily modified for each year of participation in the TRECVid conference. As such, we began by describing our interface, which is based on a modern smart device, that of the Apple iPad. Next we discussed the back-end components such as the image similarity, clustering and text search indexes and how they supported the system. We concluded this chapter by outlining how this system supports each of the experiments we ran over the course of this research.

In the first of our experiment chapters, we focused on work carried out for the TRECVid 2010 known item search task. We configured the system described in Chapter 3 to utilize two primary search methods in the form of text and classifier search and one secondary search in the form of similarity search. We had participation in the interactive experiment from two user groups, that of experts and novices. With our tailored interface, we believed that novice users would perform to a similar level of the experienced expert users. From the results attained through participation in TRECVid, both user groups performed almost identically with respect to the framework's evaluation criteria. We also did post-TRECVid analysis, with respect to number of searches performed and again both our user groups performed similarly. In this chapter, we show, that given the circumstances set out in this experiment, we have proven the first hypothesis.

For our second experimentation chapter, we focused on the TRECVid 2011 known item task and the second of our hypotheses. From the pilot run in Chapter 4, we discovered insights into novice user behavior. These users were reluctant to implement classifiers from a primary search, due to greater familiarity with text search. Post analysis showed that in certain cases that using classifiers could improve the rank of the users query. We believed that if we automatically apply the classifiers, novice users will increase their chance of attaining the known item. Another area novice users neglected was that of similarity search. To this end we implemented a clustering algorithm to group content. We believed this clustering would speed up user searches allowing for quick acceptance or dismissal of content. For this work, we reused the system outlined in Chapter 3, only using the primary searches of text and/or classification. From the results, we see the clustering system is not only faster than the non clustering system by over 30%, but it was the best performing system that year for the known item search task. The work carried out in this chapter validates the second hypothesis.

Finally, in our last experiment chapter, we conclude the three year run for TRECVid known item search task and our third hypothesis. Taking the lessons learned from both previous experiments, we try to merge aspects from both systems into a final version. From the 2010 system we adapt the multiple keyframe representation and the classifiers which were found useful from post analysis. From the 2011 system, we implement the clustering algorithm and visual keyframe selection. We devise a method for representing video in multiple clusters to allow the user a better chance at attaining the known-item. Our final experimental results show that the multi keyframe clustering system outperforms the single keyframe clustering system with respect to the TRECVid results and searches performed. We believe this satisfies the final hypothesis.

7.2 Contributions

- We implemented a system which had been tailored towards novice users. This system evolved over the course of the three year participation featuring the automatic use of visual search aids to enhance the novice users experience. With the experiments carried out in 2010 we showed that for this type of interactive experiment both novice and expert users performed similarly with respect to the chosen evaluation metrics.

- We defined a scheme which for this type of experiment gives the users an understanding of the video document based on its representation. We achieve this through not only calculating an improved keyframe representation, but also by implementing a clustering technique to group similar items in a ranked list, thereby enhancing the search experience.

- We developed the cluster-list, a new type of ranked list for content which is grouped based on visual similarity. A system which utilized the cluster-list achieved the top score in the Interactive Known-Item search experiments in TRECVid 2011.

- All of the above contributions come together to define a new way of searching for and representing video content on mobile devices, that, in our TRECVid experimentation, supported novice users to perform equally as well as expert users in known item search tasks.

7.3 Shortcomings

We can identify a number of potential shortcomings of this research that could potentially be addressed by further experimentation. One such issue is the sample size of the user experiments. While the case has been made that small heuristic experiments can be useful and insightful by Jacob Nielsen[1], to enhance significance we could require larger numbers. However, the sample size is bigger than previous generations of TRECVid, where sample sizes of four experts and a single novice are the norm. This is an area which can be addressed in future research.

Another shortcoming with the research is with respect to the chosen evaluation framework. While it does model a real-world scenario where a user is looking for a specific item, the framework is not evaluated according to metrics which show the effectiveness of the retrieval due to the single item of relevance. In the IR community, most systems are measured with metrics such as recall and accuracy towards a topic.

7.4 Future Work

Our work poses many new research questions and we will endeavour to outline where progress could be made. First, we discuss how we could improve the evaluation methodology before describing how to improve the techniques put forward by clustering, text analysis and concepts.

7.4.1 Evaluation

As stated above in the shortcomings section, our evaluation was carried out using a very small sample size and while this is fine for heuristic evaluations, a larger sample size will always be better. It would be useful to set up user experimentation with significance testing at the forefront. Larger sample sizes with appropriate user level feedback could be gathered both explicitly through user surveys and implicitly through the capture of user interaction logs with the system.

[1] http://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/

7.4.2 Clustering

While in this work we defined a successful method of clustering, in future systems we could explore options for replacing the k-means algorithm and attempt to utilize other algorithms which may perform better for video retrieval systems. We believe, that by utilizing other evidences such as text or tag clustering we can, through fusion, further improve the end-user′s experience, and greatly increase their search capabilities to that of an expert′s level without unnecessary demand on the end user.

One other suggestion is with regard to how we apply clustering. In our previous work we utilized a batch clustering approach with the clusters being predefined. This led to certain clusters containing only a few items. Future systems could implement result clustering which we believe will allow for more of the canvas to be used.

7.4.3 Text Analysis

One area with which we could improve our work is with the automatic application of concepts. We showed in Chapter 4 that novice users had a preference for text searching. Their reliance on text could be harnessed to facilitate a better automatic concept application. Through the use of text analysis we could map user queries to appropriate concepts. In this way, we could reduce the user input to a single interactive search box. We believe this will be a preferential method for novice users as it will remove complete complexity in the system allowing users to search without having to understand the use of concepts.

7.4.4 Future Content Based Retrieval

Another area of future work could be with the simplification of interaction in content based systems. We have shown that our system, contrary to traditional

CBMIR, performs as well with fewer features available to the end user. Further extension could be achieved through further analysis of automated techniques to reduce the user overhead even further by abstracting away more of the complexity.

7.4.5 Concepts

In this work, we employed a limited number of concepts. This was due to the limitations in resources available. One area in which we could improve the system's performance is by employing additional concepts which could aid the user in searching. These concepts could be identified through user query analysis and thus linked with the text analysis explained above.

7.5 Final Conclusion

In this thesis, we presented a system which participated in multiple instances of TRECVid through the years of 2010 - 2012. The system built upon experience gained through previous participation in the conference. We extended the state-of-the-art by reducing the levels of user interaction required as a result of the incorporation of content-based retrieval techniques, while at the same time reducing the complexity of the search mechanisms. So as to constrain our research to real-world uses, we have focused our development and evaluation on supporting the user in accessing content from mobile devices. In the case of this research we have specifically chosen the iPad, though our contributions could also be applied to other mobile devices, or even be used to enhance the user interaction on desktop or other non-mobile devices. Through the three years we have shown, that using classifiers can help, though users require training and their expectations to be managed. We found that removing the complexity of the system and allowing the users to search rather than focusing on hard to formulate queries gave us better results when compared with peer research groups. Overall by implementing these enhancements, we have developed a method which has proven effective in increasing users' search performance.

Appendix A

User Survey

In this section we give an example of the user documentation used in our experiments, featuring

- Instruction Sheet
- Entry Questionnaire
- Task Questionnaire
- Exit Questionnaire

Examples of such are available below.

INFORMATION SHEET

Project: **A Study of Clustering Techniques to Improve Video Search**

Researchers: **Frank Hopfgartner, David Scott, Jinlin Guo, Cathal Gurrin, Alan F. Smeaton**

You are invited to take part in a research study. Before you decide to do so, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully. Ask me if anything is not clear or if you would like more information.

The aim of this experiment is to investigate the relative effectiveness of two different multimedia search systems. We cannot determine the value of search systems unless we ask those people who are likely to be using them, which is why we need to run experiments like these. Please remember that it is the systems, not you, that are being evaluated.

It is up to you to decide whether or not to take part. If you decide to take part you will be given this information sheet to keep and asked to sign a consent form. You are free to withdraw at any time without giving a reason. You also have the right to withdraw retrospectively any consent given, and to require that any data gathered on you be destroyed.

The experiment will last around two hours. You will be given a chance to learn how to use the two systems before we begin. At this time you will also be asked to complete an introductory questionnaire. You will perform twelve tasks in total. Each task should take around 5 minutes to complete. After using each system you will be asked to fill in a questionnaire and your interactions (e.g. mouse clicks and key presses) will also be logged. You are encouraged to comment on each interface as you use it, which I will take notes on. Please ask questions if you need to and please let me know about your experience during the search. Finally, after completing all tasks, you will be asked some questions about the tasks, your search strategy and the systems. Remember, you can opt out at any time during the experiment.

All information collected about you during the course of this study will be kept strictly confidential. You will be identified by an ID number and all information about you will have your name and contact details removed so that you cannot be recognised from it. Data will be stored for analysis, and then destroyed.

The results of this study may be used for some PhD research. The results are likely to be published in late 2012. You can request a summary of the results in the consent form. You will not be identified in any report or publication that arises from this work.

This study is being funded by the Norwegian Research Council via the Information Access Disruptions (iAD) Project and Science Foundation Ireland under grant (07/CE/I1147).

For further information about this study please contact:

Dr. Frank Hopfgartner
CLARITY: Centre for Sensor Web Technologies, Dublin City University
Dublin 9, Ireland
Email: frank.hopfgartner@computing.dcu.ie
Tel.: 1 700 8563

ENTRY QUESTIONNAIRE

This questionnaire will provide us with background information that will help us analyse the answers you give in later stages of this experiment. You are not obliged to answer a question, if you feel it is too personal.

User ID:

Please place a TICK ☑ in the square that best matches your opinion.

Part 1: PERSONAL DETAILS

This information is kept completely confidential and no information is stored on computer media that could identify you as a person.

1. Please provide your AGE:

2. Please indicate your GENDER:

Male.. □ 1 Female.. □ 2

3. Please provide your current OCCUPATION: YEAR:

4. What is your FIELD of work or study?

5. What is your educational level

Undergraduate/No Degree............ □ 1	Graduate Student/Primary Degree. □ 2
Researcher/Advanced Degree....... □ 3	Faculty/Research Staff....................... □ 4

Part 2: SEARCH EXPERIENCE

Experience with Multimedia

Circle the number closest to your experience.

How often do you…	Never	Once or twice a year	Once or twice a month	Once or twice a week	Once or twice a day	More often
6. deal with videos, photographs or images in your work, study or spare time?	1	2	3	4	5	6
7. take videos or photographs in your work, study or spare time?	1	2	3	4	5	6
8. carry out image or video searches at home or work?	1	2	3	4	5	6
9. follow daily news broadcasts?	1	2	3	4	5	6
10. watch news videos online?	1	2	3	4	5	6

Multimedia Search Experience

11. Please indicate which online search services you use to search for MULTIMEDIA (mark AS MANY as apply)

Google (http://www.google.com)... □ 1

Yahoo (http://www.yahoo.com)... □ 2

AltaVista (http://www.altavista.com)... □ 3

AlltheWeb (http://www.alltheweb.com).. □ 4

YouTube (http://www.youtube.com)... □ 5

Flickr (http://www.flickr.com).. □ 6

Microsoft (http://www.live.com)... □ 7

Others (please specify)...... □

12. Using the MULTIMEDIA search services you chose in question 11 is GENERALLY:

easy	□	□	□	□	□	difficult	N/A
stressful	□	□	□	□	□	relaxing	□
simple	□	□	□	□	□	complex	
satisfying	□	□	□	□	□	frustrating	

13. You find what you are searching for on any kind of MULTIMEDIA search service…

Never				Always	N/A
□	□	□	□	□	□
1	2	3	4	5	

14. Please indicate which systems you use to MANAGE your MULTIMEDIA (mark AS MANY as apply)

None (I just create directories and files on my computer)...................... □ 1

Adobe Album……………………………….. □ 2

Picasa (Google)……………………………….. □ 3

iView Multimedia (Mac)……………………….. □ 4

ACDSee……………………….. □ 5

Others (please specify)...... □

15. Using the multimedia management tools you chose in question 14 is GENERALLY:

easy	□	□	□	□	□	difficult	N/A
stressful	□	□	□	□	□	relaxing	□
simple	□	□	□	□	□	complex	
satisfying	□	□	□	□	□	frustrating	

16. It is easy to find a particular image that you have saved previously on your computer…

Never				Always	N/A
□	□	□	□	□	□
1	2	3	4	5	

17. What do you expect from a multimedia search service?

18. What sort of features would you expect in such a multimedia search service?

POST-TASKS QUESTIONNAIRE

To evaluate the system you have just used, we now ask you to answer some questions about it. Take into account that we are interested in knowing your opinion: answer questions freely, and consider there are no right or wrong answers. Please remember that we are evaluating the system you have just used and not you.

User ID:		System:	

Please place a TICK ☑ in the square that best matches your opinion. Please answer all questions.

Part 1: TASKS

In this section we ask about the search tasks you have just attempted.

1.1. The tasks we asked you to perform were:

unclear	□	□	□	□	□	clear
easy	□	□	□	□	□	difficult
simple	□	□	□	□	□	complex
unfamiliar	□	□	□	□	□	familiar

1.2. It was easy to formulate queries on these topics.

Agree				Disagree
□	□	□	□	□
5	4	3	2	1

1.3. The search I have just performed was.

stressful	□	□	□	□	□	relaxing
interesting	□	□	□	□	□	boring
tiring	□	□	□	□	□	restful
easy	□	□	□	□	□	difficult

Part 2: RETRIEVED VIDEOS

In this section we ask you about the videos you found/selected.

2.1. The videos I have received through the searches were:

relevant	□	□	□	□	□	not relevant
inappropriate	□	□	□	□	□	appropriate
complete	□	□	□	□	□	incomplete

2.2. I had an idea of which kind of videos were relevant for the topic before starting the search.

Not at all — Vague — Clear

□	□	□	□	□
1	2	3	4	5

2.3. During the search I have discovered more aspects of the topic than initially anticipated.

Disagree — Agree

□	□	□	□	□
1	2	3	4	5

2.4. The video(s) I chose in the end match what I had in mind before starting the search.

Exactly — Not at all

□	□	□	□	□
5	4	3	2	1

2.5. I believe I have seen all possible videos that satisfy my requirement.

Agree — Disagree

□	□	□	□	□
5	4	3	2	1

2.6. I am satisfied with my search results.

Very — Not at all

□	□	□	□	□
5	4	3	2	1

Part 3: SYSTEM & INTERACTION

In this section we ask you some general questions about the system you have just used.

3.1. Overall reaction to the system:

terrible	□	□	□	□	□	wonderful
satisfying	□	□	□	□	□	frustrating
dull	□	□	□	□	□	stimulating
easy	□	□	□	□	□	difficult
rigid	□	□	□	□	□	flexible
efficient	□	□	□	□	□	inefficient
novel	□	□	□	□	□	standard
effective	□	□	□	□	□	ineffective

Not at all

3.2. When interacting with the system, I felt:

in control	□	□	□	□	□	not in control
uncomfortable	□	□	□	□	□	comfortable
confident	□	□	□	□	□	unconfident

3.3. How easy was it to LEARN TO USE the system?

Not at all

□	□	□	□	□
1	2	3	4	5

3.4. Did you find that the length of the training session for the system you used was sufficient?

Not at all

□	□	□	□	□
1	2	3	4	5

3.5. How easy was it to USE the system?

□	□	□	□	□
5	4	3	2	1

3.6. Did you find that the system response time was fast enough?

□	□	□	□	□
5	4	3	2	1

Disagree

Part 4: SYSTEM SUPPORT & SEARCH STRATEGY

In this section we ask you more detailed questions about the system and your search strategy.

4.1. The system was effective for solving the task.

□	□	□	□	□
5	4	3	2	1

Because it helped me to…	Disagree				
4.2. analyse the task.	1	2	3	4	5
4.3. explore the collection.	1	2	3	4	5
4.4. find relevant videos.	1	2	3	4	5
4.5. organise the videos I found for the task.	1	2	3	4	5
4.6. detect and express different aspects of the task.	1	2	3	4	5

4.7. How you conveyed relevance to the system (i.e. ticking boxes) was:

difficult	□	□	□	□	□	easy
effective	□	□	□	□	□	ineffective
not useful	□	□	□	□	□	useful

4.8. What was the most useful tool to support your search strategy?

4.9. What was the least useful tool to support your search strategy?

Disagree

4.10.	Do you have any other comments on the system?
e.g.	a) Did selecting images usually improve the results? b) What could be improved?

And finally:

4.11. I believe I have succeeded in my performance of the task.

Disagree

□ □ □ □ □

1 2 3 4 5

What are the issues/problems that affected your performance?					
4.12. I didn't understand the task.	1	2	3	4	5
4.13. I video collection didn't contain the video I wanted.	1	2	3	4	5
4.14. The system didn't return relevant videos.	1	2	3	4	5
4.15. I didn't have enough time to do an effective search.	1	2	3	4	5
4.16. I was often unsure of what action to take next.	1	2	3	4	5

EXIT QUESTIONNAIRE/INTERVIEW

The aim of this experiment was to investigate the relative effectiveness of two different video search systems. Please consider the entire search experience that you just had when you respond to the following questions.

User ID:

Please place a TICK ☑ in the square that best matches your opinion. Please answer the questions as fully as you feel able to.

Part 1: TASKS and SEARCH STRATEGY

1.1. To what extent did you find the tasks similar to other searching tasks you typically perform?

Not at all — Completely

□	□	□	□	□
1	2	3	4	5

1.2. How did the search tasks fit into your normal multimedia search tasks?

a) What sort of multimedia search tasks do you need to perform?
b) What sort of search tasks do you perform in order to fulfil your needs?

1.3. Describe your natural search strategy (taking a typical search task into consideration)?

a) Your problem solving strategy?
b) Is it dependent on the search task?
c) In an ideal scenario, how could a system support your search strategy?

1.4. Which of the two systems supported your strategy better?

a) How?
b) Why?
c) What did you have to do in each case to adapt your search strategy to the system?
d) In an ideal scenario (when you have the necessary tools) would you be following the same search strategy?

Part 2: TASKS and INFORMATION NEED DEVELOPMENT

2.1. How clear did you find the tasks and how well-defined was your initial information need?

Task	Unclear				Clear	**IN**	Not at all				Completely
	1	2	3	4	5		1	2	3	4	5
Task 1	□	□	□	□	□	Task 1	□	□	□	□	□
Task 2	□	□	□	□	□	Task 3	□	□	□	□	□
Task 3	□	□	□	□	□	Task 2	□	□	□	□	□
Task 4	□	□	□	□	□	Task 4	□	□	□	□	□

2.2. How did your need develop?

a) Did you get new ideas discover or new aspects of the task during the search?
b) What caused you to change your initial idea?
c) How did the system support/trigger changes?
d) Which of the systems was more helpful in developing your information need?

PART 3: SYSTEM EXPERIENCE

Which of the systems did you…	System 1	System 2	No difference
3.1. … find easier to LEARN TO USE?	□	□	□
3.2. … find easier to USE?	□	□	□
3.3. … find more EFFECTIVE for the tasks you performed?	□	□	□
3.4. … LIKE BEST overall?	□	□	□

3.5. What did you LIKE about each of the systems?

System 1:

System 2:

3.6. What did you DISLIKE about each of the systems?

System 1:

System 2:

Appendix B

Known-Item Search Topic Descriptions

The below are topic descriptions taken from the known-item search task of TRECVid 2012.

0891 1-5 KEY VISUAL CUES: geysers, bus, flags

0891 QUERY: Find a video of yellow bus driving down winding road in front of building with flags on roof and driving past geysers

0892 1-5 KEY VISUAL CUES: lake, trees, boats, buildings

0892 QUERY: Find the video with panned scenes of a lake, tree-lined shoreline and dock with several boats and buildings in the background.

0893 1-5 KEY VISUAL CUES: man, soccer ball, long hair, green jacket, parking lot, German

0893 QUERY: Find the video of man speaking German with long hair and green jacket and soccer ball in a parking lot.

0894 1-5 KEY VISUAL CUES: Russian jet fighter, red star, white nose cone, sky rolls, burning airship

0894 QUERY: Find the video of an advance Russian jet fighter with red star on wings and tail and a white nose cone that does rolls in the sky and depicts a burning airship

0895 1-5 KEY VISUAL CUES: Sunday Quickie, man with glasses in blue shirt standing by window, raining, wooden fence

0895 QUERY: Find the video titled ”Sunday Quickie” of a man who is wearing glasses and a blue shirt standing by the window and watching the rain outside and discussing his trip to Home Depot and Harveys Hamburger Kiosk.

0896 1-5 KEY VISUAL CUES: Yucca mountain

0896 QUERY: Find the video including a shot of Yucca Mountain.

0897 1-5 KEY VISUAL CUES: strobe lights, yarn, men

0897 QUERY: Find the video of strobe lights with men wrapped up in yarn

0898 1-5 KEY VISUAL CUES: man-gray T shirt, man-white hat, yellow disc, man-tree, ski/equipment lift

0898 QUERY: Find the video depicting a man wearing a gray T shirt and a man wearing a white hat tossing a yellow disc with one man climbing a tree and ski/equipment lift overhead.

0899 1-5 KEY VISUAL CUES: man, airplane, glasses, beard

0899 QUERY: Find the video with a man with glasses, a red jacket and black shirt, white hair and beard talking in front of a Navy airplane.

0900 1-5 KEY VISUAL CUES: people, street, celebration

0900 QUERY: Find the video with people celebrating in a street yelling and running

0901 1-5 KEY VISUAL CUES: paper, tree, animation

0901 QUERY: Find the video with no sound showing a white paper with marks followed by an animation drawing of a tree.

0902 1-5 KEY VISUAL CUES: caskets, American flag

0902 QUERY: Find the video with several rows of caskets, each draped in an American flag.

0903 1-5 KEY VISUAL CUES: left side of rainbow, mountains, shrubbery

0903 QUERY: Find the video with the left side of a rainbow. It is dark outside and there are mountains in the background to the right and shrubbery in the foreground.

0904 1-5 KEY VISUAL CUES: head shots, women, men, conference, podcamp

0904 QUERY: Find the video with a head shots of men and women at a conference outside of Boston talking about their experiences at a podcamp

0905 1-5 KEY VISUAL CUES: Sigma Alien 2 advertisement, red car, man/light, helicopter, aliens, greenish gun fire

0905 QUERY: Find the video of a Sigma Alien 2 advertisement where a man shines a light on a red car while dropping down from an helicopter into a building complex and immediately coming under alien gun attack and returning fire which appears in greenish color.

0906 1-5 KEY VISUAL CUES: man-brown hair, brown Tshirt, Brown chair, animated speak, Obama's inauguration parties

0906 QUERY: Find the video of a man with brown hair wearing a brown Tshirt sitting in a brown chair animatedly speaking against Obama's inauguration parties.

0907 1-5 KEY VISUAL CUES: Six steps down and into the mud, Blur of images with some images of people visible, Sliabh Cairn

0907 QUERY: Find the video that contains the sentence "Six steps down and into the mud" on the screen in the beginning followed by blurred images, people visible in one portion, and the name "Sliabh Cairn" appearing on screen near the end.

0908 1-5 KEY VISUAL CUES: orchestra, stage, chorus, cartoon, horse

0908 QUERY: Find the video showing an orchestra playing on a stage with a chorus standing behind. Also, a cartoon character rides a horse.

0909 1-5 KEY VISUAL CUES: woman, black oil, milk carton

0909 QUERY: Find the video of woman pouring black oil from milk carton.

0910 1-5 KEY VISUAL CUES: wall, garden, camera, picture frame, bed

0910 QUERY: Find the video of a man taking pictures of people in a walled garden. A woman stands on a tree stump higher than two men, creating a pyramid shape. Inside on a bed the three look through a picture frame. The film is blurry and distorted.

0911 1-5 KEY VISUAL CUES: young man, sofa, notebook, text, two purple stripes.

0911 QUERY: Find the video with two purple vertical stripes on the left side of the screen. There is text relating to the murder of a man and his son's swearing to get revenge against his father's murderer. A young man with glasses and dark brown hair is sitting on a sofa holding a notebook and pen.

0912 1-5 KEY VISUAL CUES: bathroom, brown walls, checked curtains, camel

0912 QUERY: Find the video of bathroom with brown walls, checked curtains and picture of camel on wall.

0913 1-5 KEY VISUAL CUES: web address of www.avezpasa.co.yu, older man with jacket, old woman in black with scarf, dog

0913 QUERY: Find the video in a foreign language with the url - www.avezpasa.co.yu showing throughout and of a man interviewing different people in a town some of which includes: an old scarfed sitting woman in black and older man in a suit jacket standing next to stores. A picture of dog at end of clip.

0914 1-5 KEY VISUAL CUES: Secretary Rice, bald man, committee meeting

0914 QUERY: Find the video at a committee meeting on capitol hill with Secretary Rice speaking and a bald headed man protesting and being escorted out

Appendix C

User Log Examples

The following is an example of the interaction logs with the system for a single topic where the user sucessfully finds the known item. The topic in question is to find a video with a demonstration of the Sega game featuring tanks called hounds. The query is formulated by a text query followed by 34 zeros which identify classifiers issued by the user.

```
===================INITIALISATION=========================
User : 3
Topic : 2
Timestamp : 17:17 09/09/2010
=========================================================
=====================QUERY SECTION========================
Query : sega hounds0000000000000000000000000000000000
Timestamp : 17:17 09/09/2010
=========================================================
===============SHOT TIMING SECTION=========================
Shot : 8095_1
Timestamp : 17:17 09/09/2010
=========================================================
=====================QUERY SECTION========================
Query : sega hounds advertisement0000000000000000000000000000000000
Timestamp : 17:18 09/09/2010
==========================================================
===============SHOT TIMING SECTION========================
Shot : 6422_1
Timestamp : 17:19 09/09/2010
==========================================================
=====================QUERY SECTION========================
Query : sega advertisement0000000000000000000000000000000000
Timestamp : 17:19 09/09/2010
==========================================================
===============SHOT TIMING SECTION========================
Shot : 4401_1
Timestamp : 17:19 09/09/2010
==========================================================
===================VALIDATION SECTION======================
Video : 4401
time elapsed : 146
verdict: true
Timestamp : 17:19 09/09/2010
==========================================================
```

Bibliography

Auffret, G., Foote, J., Li, C.-S., Shahraray, B., Syeda-Mahmood, T., and Zhang, H. (1999). Multimedia access and retrieval (panel session): the state of the art and future directions. In Chang, S.-F., editor, *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, MULTIMEDIA '99, pages 443–445, New York, NY, USA. ACM.

Ayache, S. and Quenot, G. (2008). Video Corpus Annotation using Active Learning. In *European Conference on Information Retrieval (ECIR)*, pages 187–198, Glasgow, Scotland.

Baeza-Yates, R. A. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Blanken, H., de Vries, A., Blok, H., and Feng, L., editors (2007). *Multimedia Retrieval*. Data-Centric Systems and Applications. Springer Verlag, Berlin.

Browne, P., Gurrin, C., Lee, H., Donald, K. M., Sav, S., Smeaton, A. F., and Ye., J. (2001). Dublin city university video track experiments for trec 2001. In *TREC 2001 - Text REtrieval Conference*, MD, USA. National Institute of Standards and Technology.

Browne, P., Smeaton, A. F., Murphy, N., O'Connor, N., Marlow, S., and C., B. (2000). Evaluating and combining digital video shot boundary detection algorithms. In *IMVIP 2000 - Irish Machine Vision and Image Processing Conference*, pages 93–100.

Cai, D., He, X., Li, Z., Ma, W.-Y., and Wen, J.-R. (2004). Hierarchical clustering of www image search results using visual, textual and link information. In *Proceedings of the 12th annual ACM international conference on Multimedia*, MULTIMEDIA '04, pages 952–959, New York, NY, USA. ACM.

Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y., and Moon, S. (2007). I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, IMC '07, pages 1–14, New York, NY, USA. ACM.

Chaisorn, L., Zheng, Y.-T., and Sim, K. (2011). Known-item search (kis) in video: Survey, experience and trend. In *Information, Communications and Signal Processing (ICICS) 2011 8th International Conference on*, pages 1 –4.

Chen, C., Gagaudakis, G., and Rosin, P. L. (2000). Content-based image visualization. In *IV*, pages 13–18. IEEE Computer Society.

Chen, M., Li, H., and Hauptmann, A. (2009). Informedia@ trecvid 2009: Analyzing video motions. In *In Proceedings of the 7th TRECVID Workshop, Gaithersburg, USA, November 2009.*

Christel, M. G. and Conescu, R. M. (2006). Mining novice user activity with trecvid interactive retrieval tasks. In *Proceedings of the 5th international conference on Image and Video Retrieval*, CIVR'06, pages 21–30, Berlin, Heidelberg. Springer-Verlag.

Christel, M. G., Hauptmann, A. G., Hauptmann, E. G., Warmack, A. S., and Crosby, S. A. (1999). Adjustable filmstrips and skims as abstractions for a digital video library. In *IEEE Advances in Digital Libraries Conference*, pages 98–104. IEEE Press.

Cleverdon, C. (1997). Readings in information retrieval. chapter The Cranfield tests on index language devices, pages 47–59. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Dunlop, M. and Brewster, S. (2002). The challenge of mobile devices for human computer interaction. *Personal Ubiquitous Comput.*, 6(4):235–236.

Ebner, M., Stickel, C., and Kolbitsch, J. (2010). iphone/ipad human interface design. In Leitner, G., Hitz, M., and Holzinger, A., editors, *HCI in Work and Learning, Life and Leisure*, volume 6389 of *Lecture Notes in Computer Science*, pages 489–492. Springer Berlin Heidelberg.

Ebrahimi, T. and Horne, C. (2000). C.: Mpeg-4 natural video coding an overview. In *Signal Processing: Image Communication 15*, pages 365–385.

Foley, C., Guo, J., Scott, D., Ferguson, P., Wilkins, P., McCusker, K., Diaz, E. S., Gurrin, C., Smeaton, A. F., i Niero, X. G., Marques, F., McGuinness, K., and O'Connor., N. E. (2010). Trecvid 2010 experiments at dublin city university. In *TRECVid 2010 - Text REtrieval Conference TRECVid Workshop*.

Foley, C., Gurrin, C., Jones, G., Lee, H., Givney, S. M., OConnor, N., Sav, S., Smeaton, A. F., and Wilkins, P. (2005). Trecvid 2005 experiments at dublin city university. In *TRECVid 2005 - Text REtrieval Conference TRECVid Workshop*.

Foley, C. and Smeaton., A. F. (2010). Division of labour and sharing of knowledge for synchronous collaborative information retrieval. *Information Processing and Management*, 46(6):762–772.

Gauvain, J.-L., Lamel, L., and Adda, G. (2002). The limsi broadcast news transcription system. *Speech Commun.*, 37(1-2):89–108.

Gurrin, C., Brenna, L., Zagorodnov, D., Lee, H., Smeaton, A. F., and Johansen, D. (2006). Supporting mobile access to digital video archives without user queries.

In *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services*, MobileHCI '06, pages 165–168, New York, NY, USA. ACM.

Gurrin, C., Lee, H., Caprani, N., Zhang, Z., OConnor, N., and Carthy, D. (2010). Browsing large personal multimedia archives in a lean-back environment. In Boll, S., Tian, Q., Zhang, L., Zhang, Z., and Chen, Y.-P., editors, *Advances in Multimedia Modeling*, volume 5916 of *Lecture Notes in Computer Science*, pages 98–109. Springer Berlin Heidelberg.

Hanjalic, A. (2002). Shot-boundary detection: unraveled and resolved? *Circuits and Systems for Video Technology, IEEE Transactions on*, 12(2):90 –105.

Hauptmann, A. G. (2004). Successful approaches in the trec video retrieval evaluations. In *In Proc. ACM Multimedia*, pages 668–675. ACM Press.

Hauptmann, E. G., hao Lin, W., Yan, R., Yang, J., and yu Chen, M. (2006). Extreme video retrieval: joint maximization of human and computer performance. In *In ACM Multimedia*, pages 385–394. ACM Press.

Heng, W. J. and Ngan, K. N. (2002). Shot boundary refinement for long transition in digital video sequence. *Multimedia, IEEE Transactions on*, 4(4):434 – 445.

Hürst, W. and Meier, K. (2008). Interfaces for timeline-based mobile video browsing. In *Proceedings of the 16th ACM international conference on Multimedia*, MM '08, pages 469–478, New York, NY, USA. ACM.

Hürst, W., Snoek, C. G. M., Spoel, W. J., and Tomin, M. (2010). Keep moving! revisiting thumbnails for mobile video retrieval. In *ACM International Conference on Multimedia*.

Hürst, W., Snoek, C. G. M., Spoel, W. J., and Tomin, M. (2011). Size matters! how thumbnail number, size, and motion influence mobile video retrieval. In *International Conference on MultiMedia Modeling*, pages 230–240.

Jacobs A., Miene A., I. G. T. and O., H. (2004). Automatic shot boundary detection combining color, edge, and motion features of adjacent frames. In *In TRECVID 2004*, pages pp. 197–206.

Jiang, Y.-G., Yang, J., Ngo, C.-W., and Hauptmann, A. G. (2010). Representations of keypoint-based semantic concept detection: A comprehensive study. *IEEE Transactions on Multimedia*, 12(1):42–53.

Jones, K. S. (1980). *Search term relevance weighting - some recent results*. Journal of Information Science,, Journal of Information Science, 1, 1980, 325-332.

Klaus, S., Frank, H., Oge, M., Laszlo, B., and M., J. J. (2010). Video browsing interfaces and applications: a review. *Journal of Photonics for Energy*, pages 018004–018004–35.

Lee, H., Smeaton, A., Murphy, N., OConnor, N., and Marlow, S. (2001). Handheld user interface design to a video indexing, browsing, and playback system. *In:Proc. UAHCI*.

Lee, H. and Smeaton, A. F. (2002a). Designing the user-interface for the fschlr digital video library. In *Journal of Digital Information*.

Lee, H. and Smeaton, A. F. (2002b). Searching the fschlr-news archive on a mobile device. In *Proceedings of the 25th International ACM Conference on Research and Development in Information Retrieval (SIGIR 2002), Workshop on Mobile Personal Information Retrieval*, pages 11–15.

Lei, Z., Wu, L.-D., Lao, S.-Y., Wang, G., and Wang, C. (2004). A new video retrieval approach based on clustering. In *Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on*, volume 3, pages 1733 – 1738 vol.3.

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM J. Res. Dev.*, 1(4):309–317.

Nguyen, G. P. and Worring, M. (2008). Interactive access to large image collections using similarity-based visualization. *J. Vis. Lang. Comput.*, 19(2):203–224.

O'Connor, N., Lee, H., Smeaton, A., Jones, G., Cooke, E., Le Borgne, H., and Gurrin, C. (2003). Dcu experiments @ trecvid 2003. In *Proceedings TRECVid 2003*.

O'Connor, N., Lee, H., Smeaton, A., Jones, G., Cooke, E., Le Borgne, H., and Gurrin, C. (2006). Fischlar-trecvid-2004: combined text- and image-based searching of video archives. In *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on*, page 4 pp.

Over, P., Awad, G., Fiscus, J. G., Antonishek, B., Michel, M., Kraaij, W., Smeaton, A. F., and Quénot, G. (2010). Trecvid 2010 - an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID*.

Over, P., Awad, G., Michel, M., Fiscus, J., Sanders, G., Shaw, B., Kraaij, W., Smeaton, A. F., and Quenot, G. (2012). Trecvid 2012 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2012*. NIST, USA.

Qunot, G., Moraru, D., and Besacier, L. (2003). Clips at trecvid: Shot boundary detection and feature detection. In *TRECVID 2003 Workshop Notebook Papers*, pages 35–40.

Rijsbergen, C. J. V. (1979). *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition.

Salembier, P. and Sikora, T. (2002). *Introduction to MPEG-7: Multimedia Content Description Interface*. John Wiley & Sons, Inc., New York, NY, USA.

Salton, G. (1968). *Automatic Information Organization and Retrieval.* McGraw Hill Text.

Salton, G. (1971). *The SMART Retrieval System&#8212;Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Sav, S., Lee, H., Smeaton, A., and OConnor, N. (2006). Using segmented objects in ostensive video shot retrieval. In Detyniecki, M., Jose, J., Nrnberger, A., and Rijsbergen, C., editors, *Adaptive Multimedia Retrieval: User, Context, and Feedback*, volume 3877 of *Lecture Notes in Computer Science*, pages 155–167. Springer Berlin Heidelberg.

Shaw, J. A., Fox, E. A., Shaw, J. A., and Fox, E. A. (1994). Combination of multiple searches. In *The Second Text REtrieval Conference (TREC-2*, pages 243–252.

Smeaton, A. F., Foley, C., Gurrin, C., Lee, H., and McGivney, S. (2006a). Collaborative searching for video using the f&#237;schl&#225;r system and a diamondtouch table. In *Proceedings of the First IEEE International Workshop on Horizontal Interactive Human-Computer Systems*, TABLETOP '06, pages 151–159, Washington, DC, USA. IEEE Computer Society.

Smeaton, A. F., Lee, H., O'Connor, N. E., Marlow, S., and Murphy, N. (2003). Tv news story segmentation, personalisation and recommendation.

Smeaton, A. F., Over, P., and Doherty, A. R. (2010). Video shot boundary detection: Seven years of trecvid activity. *Comput. Vis. Image Underst.*, 114(4):411–418.

Smeaton, A. F., Over, P., and Kraaij, W. (2006b). Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA. ACM Press.

Smeaton, A. F. and Quigley, I. (1996). Experiments on using semantic distances between words in image caption retrieval. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, pages 174–180, New York, NY, USA. ACM.

Smeulders, A., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12):1349 –1380.

Snoek, C. G. M. and Worring, M. (2009). Concept-based video retrieval. *Found. Trends Inf. Retr.*, 2(4):215–322.

Snoek, C. G. M., Worring, M., Koelma, D. C., and Smeulders, A. W. M. (2007). A learned lexicon-driven paradigm for interactive video retrieval. *IEEE Trans. Multimedia*, 9:280–292.

Spence, R. (2002). Rapid, serial and visual: a presentation technique with potential. *Information Visualization*, 1(1):13–19.

Tong, S. and Chang, E. (2001). Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia*, MULTIMEDIA '01, pages 107–118, New York, NY, USA. ACM.

Tsai, F., Etoh, M., Xie, X., Lee, W.-C., and Yang, Q. (2010). Introduction to mobile information retrieval. *Intelligent Systems, IEEE*, 25(1):11–15.

Vakali, A., Hacid, M.-S., and Elmagarmid, A. K. (2004). Mpeg-7 based description schemes for multi-level video content classification. *Image Vision Comput.*, 22(5):367–378.

van de Sande, K. E. A., Gevers, T., and Snoek, C. G. M. (2010). Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596.

van de Sande, K. E. A., Gevers, T., and Snoek, C. G. M. (2011). Empowering visual categorization with the gpu. *IEEE Transactions on Multimedia*, 13(1):60–70.

Vogt, C. C. (2000). How much more is better? - characterizing the effects of adding more ir systems to a combination. In *In Content-Based Multimedia Information Access (RIAO*, pages 457–475.

Wilkins, P. (2007). Automatic query-time generation of retrieval expert coefficients for multimedia retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 924–924, New York, NY, USA. ACM.

Wilkins, P., Adamek, T., Jones, G., O'Connor, N., and Smeaton., A. F. (2007). Trecvid 2007 experiments at dublin city university. In *TRECVid 2007 - Text REtrieval Conference TRECVid Workshop*.

Worring, M., Snoek, C., de Rooij, O., Nguyen, G., and Smeulders, A. (2007). The mediamill semantic video search engine. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–1213 –IV–1216.

Xie, X., Lu, L., Jia, M., Li, H., Seide, F., and Ma, W.-Y. (2008). Mobile search with multimodal queries. *Proceedings of the IEEE*, 96(4):589–601.

Zhang, H., Kankanhalli, A., and Smoliar, S. W. (1993). Automatic partitioning of full-motion video. *Multimedia Syst.*, 1(1):10–28.

Zheng, W., Yuan, J., Wang, H., Lin, F., and Zhang, B. (2005). A novel shot boundary detection framework. In *in Proc. SPIE Vis. Commun. Image Process*, pages 410–420.