Interactive Video Retrieval

Huang, Zheng

Department of Computing Science

Faculty of Computing Science, Mathematics and Statistics

University of Glasgow

UNIVERSITY
of
GLASGOW

Submitted for Degree of Msc By Research

at the University of Glasgow

October 2005

To my parents.

Acknowledgements

While it is not possible to acknowledge everyone who was of assistance during the development of this dissertation, several individuals merit a special word of appreciation as follows.

In particular, I would like to pay tribute to my supervisor, Dr. Joemon Jose, who had enough faith in my abilities to give me this chance to finish Msc by research Degree, for his invaluable help, encouragement, guidance in completing this dissertation and generous availability. My appreciations also go to Keith van Rijsbergen, Programme Leader, for his advice in the beginning of my degree.

I would like to thank members of the Information Retrieval Group, past and present, for making the trip a pleasant one. Thank you all for your friendship, support and interest in my work. The administration and support staff in the Department of Computing Science deserve a huge mention for keeping everything running smoothly.

Furthermore, I dedicate this MSC by research degree to my parents, Changkai Huang, Lichuan Ruan, my elder sister--Wei Huang, for giving me the sacrifice, understanding, pursuing, great support and endless love, without which the completion of the MSC By Research would be a mission impossible.

Abstract

Video storage, analysis, and retrieval has become an important research topic recently due to the advancements in the creation and distribution of video data. In this thesis, an investigation into interactive video retrieval is presented.

Advanced feedback techniques have been investigated in the retrieval of textual data. Novel interactive schemes, mainly based on the concept of relevance feedback, have been developed and experimented. However, such approaches have not been applied in the video retrieval domain.

In this thesis, we investigate the use of advanced interactive retrieval schemes for the retrieval of video data. To understand the role of various features for the video retrieval, we experimented with various retrieval strategies. We benchmarked the role of visual features, the textual features and their combination. To explore this further, we categorized query into various classes and investigated the retrieval effectiveness of various features and their combination.

Based on the results, we developed a retrieval scheme for video retrieval. We developed an interactive retrieval technique based on the concept of implicit feedback. A number of retrieval models are developed based on this concept and benchmarked with a simulation-based evaluation strategy. A Binary Voting Model performed well and has been reformed for user-based experiments. We experimented with the users and compared the performance of an interactive retrieval system, using a combination of implicit and explicit feedback techniques, with that of a system using explicit feedback techniques.

Table of Content

Figure and Table List

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 1

Introduction

With the development of computer technology, digital video is rapidly becoming the medium of choice for entertainment, education, and communication, and in addition much of the footage being produced could potentially be of historical and cultural importance. There is massively increasing demand of both professional and amateur digital video.

According to the sales report of Canon (Canon, 2003), sales of digital cameras greatly increased by 1041% between 2000 and 2003, are gradually starting to replace the products of traditional analogue formats. The low cost of high speed IDE hard drives, also makes available the large amounts of storage necessary for large video files. 64-bit processors like the Apple G5, AMD's 64bit technique, Intel 64bit CPU, and multi-core processors technique greatly reduce the long latency traditionally associated the computationally intensive video editing, compression and retrieval. All of these contribute to video being one of the areas where technology is opening up huge possibilities for future more usage of video content - video clip databases for broadcast companies, video editing systems for film producers and various home video entertainment systems such as DVD, Web TV. TiVo (TiVo,2003) and (ReplayTV,2005), both of which are currently the digital version of the conventional VCR, overcome the problems inherent in conventional systems, such as degrading video quality and managing a number of analogue videos, recording broadcast TV programmes in a digital format on their internal disks. Due to the technological advances and the availability of cheap and powerful hardware, large volumes of multimedia data have been created and accumulated such as Open Video Project, Informedia Project.

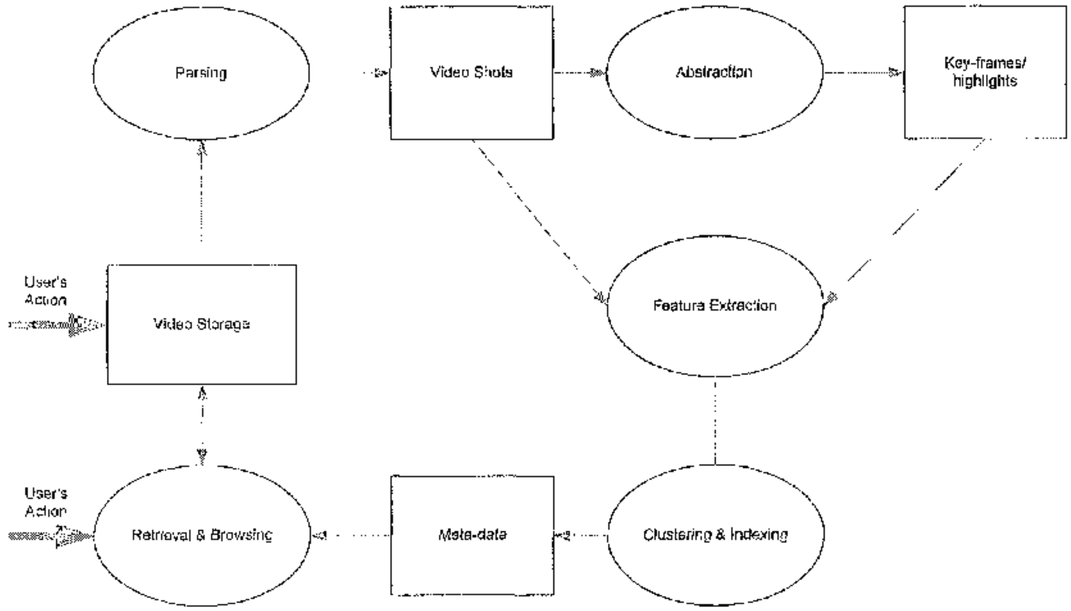
1.1 Video Retrieval

With the development of multimedia technologies, which provide comprehensive and intuitive information for a broad range of applications (Feng, 2003), videos are being digitized and made available through various information systems and/or the WWW. The digitalization of more and more videos results in a significant increase in demands for video resources and querying for video retrieval becoming more prevalent in everyday information seeking (Spink, 2001). Browne (2001) summarized "Multimedia information retrieval has significantly evolved over recent years with the development of many digital libraries and the WWW allowing browsing and retrieval of multimedia content." As a result, it is not surprising that video retrieval is becoming a very important research area.

1.1.1 Video Analysis, Browsing and Retrieval

The issues involving videos have become the most challenging research topics in various areas of multimedia technologies. It can be divided into three different areas- video analysis, browsing, and retrieval. Figure 1.1 shows the basic relationship between video analysis, browsing and retrieval (Feng, 2003).

Figure 1.1 Process Diagram for Video Content Analysis and Retrieval



1.1.1.1 Video Analysis

In general, it is simply said that video is viewed as a sequence of frames. It is more important to view video as a structured medium in which actions and events take place in time and space, comprise stories or convey particular visual information (Feng *et al.*, 2003). A video application should analyze a video as a structured document rather than a non-structured sequence of frames. For the application of video retrieval, indexing, which is the processing of creating a database of information based on the structure of video, is the work we have to do before doing retrieval, just in other kinds of retrieval (e.g. text retrieval). Video analysis can be considered as a preprocessing step for video retrieval.

1.1.1.2 Video Retrieval

Video Retrieval is the problem of searching a video document which calls for that of formulating a meaningful and clear query with the representation and similarity measures. The human searcher formulates a meaningful and clear query, the video retrieval system builds effective internal representation of features and implement similarity measures to compute similarity between user's query and a video document.There are three main types of queries—visual query, motion query, and textual query (Stéphane *et al.*, 2002).

According to Stéphane *et al.* 2002, visual query is a kind of query which uses visual objects as elements of a query because it seems natural to proceed a user's search based on examples of such visual documents, such as video documents and image documents. A query-by-example (QBE) is the main form of visual query. In a video retrieval system, key-frame (A frame selected at the beginning or end of a sequence of frames, that is used as a reference for any of a variety of functions. In inter-frame video compression, key-frames typically store complete information about the image, while the frames in between may store only the differences between two key-frames (key-frame)), motion information and a video example can be considered as visual objects. QBE-based systems have demonstrated their superior descriptive power (Zhang, 1995; Ardizzone, 1996). However, Textual keywords are the simplest way of expressing a query in a traditional IR system. For a collection of video documents, textual querying may be of even more comparative importance since it is related to high-level semantic concepts. In other words, using a textual query, the user is able to express high level concepts which would be difficult to express through QBE. Therefore, the necessity of using a combined query system appears clearly from the above reasons. However, combining query types in a video retrieval system calls for mixing parameters which may not be fully coherent with one another. Different strategies should be envisaged. One may think of using each type of query separately and combining the different results thus obtained with respect to a common relevance measure. Another simple way to combine the various querying approaches is to normalise the influence of each and to ask the users themselves to provide weights for each component of the query. This is not an acceptable solution for two major reasons (Stéphane *et al.*, 2002), which is the complication of the query formulation and simple transfer of the problem because of the underestimating of such weights.

1.1.1.3 Video Browsing

Video browsing is a special interaction mode distinguished from other kinds of retrieval. In browsing, users can obtain a new abstract representation or summary of a video. According to Feng *et al.* 2003, browsing means that an informal but quick access to content is possible. For the purpose of achieving content-based browsing, it is required to represent the information or structure of the video in a more abstract and /or summarized manner.

1.1.2 TRECVID

For the purpose of promoting the development of tools for cataloguing and retrieving digital video, the National Institute of Standards and Technology has added a video track to its TREC[1] workshop, the goal of which is to encourage research in video information retrieval by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results (TRECVID). Whether under the umbrella of TRECVID or not, there are numerous research groups working on some aspect of this problem, such as automatic video segmentation, novel descriptors of shots, or impressive user interfaces for efficient video information retrieval. The following section will introduce three projects that are based on TRECVID (TRECVID, 2003).

1.2 Examples of Video Retrieval Systems

In this section, three state-of-the-art video retrieval systems participating in the TRECVID activities will be introduced.

1.2.1 Informedia System(CMU)

The Informedia system is a digital video library which is developed by the Carnegie Mellon University for the purpose of research in the area of multimedia information

[1] http://trec.nist.gov

retrieval. This system has the following characteristics/features- "full-content search and retrieval of CNN news (1996-present), WQED[2] public broadcasts, documentaries, distinguished lectures, and other education programs". It provides a flexible multimodal query input interface which allows dynamic weight adjustments for different modalities, and integrates relevance feedback for reformulating the query and obtaining more accurate results by offering options for explicit feedback to make it available to directly control query and articulate user needs (Zhong, 2000;Hauptmann, 2003).

1.2.2 DCU system--*Físchlár*

Físchlár is a web-based digital video system that records and analyses TV broadcast programmes, developed at the Centre for Digital Video Processing in Dublin City University. It is a fully-automated system which records broadcast TV programmes on users' requests. It applies its video indexing technique--shot boundary detection, segmenting the video into individual camera shots then extracting significant key-frames from each of the camera shots. The user can then browse through the video content using several distinctive key-frame browsing interfaces, and play the recorded programme by streamed playback from a high-capacity video server. All these features of recording, indexing, browsing and playback have been integrated into a single, coherent system, running 24 hours a day on a web server (Browne *et al.*,2001;Lee, 2001).

1.2.3 Open Video Project

Open Video Digital Library is a web-based digital library, which aims to capitalize on advances in engineering as well as in library and information science to create usable services for the research and educational communities (Marchionini and Geisler, 2002). A wide range of problems, such as tests of algorithms for automatic segmentation, summarization, and creation of surrogates can be studied based on this platform.

[2] http://www.wqed.org

1.3 Relevance feedback (RF) Techniques

According to Salton and Buckley 1990, "Relevance feedback is an automatic process, introduced over, designed to produce improved query formulations following an initial retrieval operation." It is the main method of automatically reformulating the initial query for the purpose of improving a system's representation of a searcher's information need based on the feedback provided by the use in which items in initial result set actually relevant.

The technique assumes the underlying need is the same across all feedback iterations (Bates, 1989) and generally relies on explicit relevance assessments provided by the searcher (Belkin, 1996b). These indications of which documents contain relevant information are used to create a revised query that is more similar to those marked and discriminates between those marked and those not. The principal idea of RF is enhancing or weakening the importance of terms or expressions, attached to certain previously retrieved documents that have been identified as relevant by the users after doing an initial query that is the original representation of users' information need (Salton and Buckley *et al.*, 1990). The technique has been shown to be effective in non-interactive environments (Buckley, 1994), but the need to explicitly mark relevant documents is often evidence since searchers may be unwilling to directly provide relevance information. The user interface challenge is therefore to provide an easy and effective way to control the use of RF in systems that implement it. In this thesis, explicit as well as implicit techniques to gather a searcher's interests are examined.

1.3.1 Advantages of RF

The main advantages of RF are the following:

- It makes the user not have to know the details of the query formulation process, and make the construction of useful statements not depending on intimate knowledge of collection make-up and search environment.

- It makes the search session become a gradual process by separating the search operation in to several ordered steps

- It has the ability to control the query formulation process by emphasizing some terms and deemphasizing others as required in particular search case.

The major disadvantage of relevance feedback is that it increase burden on user (Xu, 1997).

In general, RF can be divided into two main types: explicit feedback, and implicit feedback; these two different feedback models will be introduced in the following chapter.

1.3.2 Explicit Feedback Model

According to White 2003, "Explicit feedback is the technique which relies on explicit relevance assessments (i.e. indications of which documents contain relevant information), and creates a revised query attuned to those documents marked." (White *et al.* 2003).

The advantage of explicit feedback is that the relevance information of documents obtained from user's explicit assessments, is clear and accurate. If explicit feedback is possible, it is a way to maximize the effectiveness of learning from information returned by a search engine (Shen, 2003). One of the disadvantages of this technique is that it is a must that users explicitly mark the relevance of documents. It means that searchers have to do more extra works, which they are reluctant to do. In addition, the relevance of a document to a search topic is often ambiguous and it is often hard for an assessor to judge precisely whether it is relevant to the topic or not because it is possible that various topics may be contained in the document a user accessed. Furthermore, the confusion on relevance of the document may make the user feel under pressure. In White 2002b, he suggested that explicit feedback can be substituted by implicit feedback, where the system attempts to estimate what the searcher may be interested in, to some extent.

1.3.3 Implicit Feedback Model

Implicit Feedback is a kind of feedback technique different from explicit feedback. The difference is the way in which the relevance judgments are obtained or inferred. in which an IR system captures search behaviour and any other interest indicator selectively and shields users from explicitly indicating which documents are relevant or non-relevant. Search behaviour is considered as implicit relevance indications, gathered from the users' interaction with the IR system.

The advantage of implicit feedback is that users do not have to explicitly mark the relevance of documents retrieved because of some cases in which it is very difficult for users to do these assessments but it is OK for the system to guess the relevance based on the searcher's behaviour. But the disadvantage of this technique is the information about the relevance of documents is not as accurate and clear as explicit feedback technique(White, 2002). However, it was also suggested that whilst not being as accurate as traditional 'explicit' RF, implicit RF (or *implicit feedback*) can be an effective substitute for its explicit counterpart in interactive information seeking environments (White *et al.*,2002b).

1.4 Research Problem

In this section, firstly I introduce the main problem addressed in this thesis and then I outline the structure of complete thesis.

1.4.1 Main Problem

The main problem in the development of effective video retrieval systems is the issue of semantic gap. Semantic gap refers to the use of low-level features for the representation of non-textual media and the failures of low-level features in associating with high-level concepts users are accustomed to. But these low-level features have a positive effect on the query formulation process (Urban and Jose, 2004). Traditionally, relevance feedback techniques are employed to address the query formulation difficulties. As discussed above, RF is the main post-query method for automatically improving a system's representation of a searcher's information need. RF can be seen as a technique to address the semantic gap issue too in Urban and Jose *et al.* 2004 and hence video retrieval systems can also benefit from the use of RF techniques.

RF techniques have not been used in practice mainly due to the cognitive issues associated with providing such feedback. The cognitive effort is too difficult for the human searcher, but it can be made by the system. Implicit feedback systems address such issues. In implicit feedback approach, the system unobtrusively monitors search behavior, and thus removes the need for the user to explicitly indicate which documents are relevant (White *et al.* 2003). For an IR system with the implicit feedback technique, the most important issue is to gather implicit relevance indications from the searcher's action for modifying the initial query. In the case, some factors of user behaviour have been most extensively investigated as sources of implicit feedback, for example, reading time, saving, printing and selecting text in the retrieval domain, which can provide implicit evidence of searcher interests (Claypool, Le, Waseda and Brown, 2001;Kelly, 2004). Although the implicit factors are generally thought to be less accurate than explicit factors, there is no extra cognitive cost for gathering large quantities of implicit data. Information about what results are relevant is obtained implicitly, by interpreting a searcher's selection of one search result over others as an indication that that result is more relevant. The Ostensive Model is also based on such principles and uses passive observational evidence, interpreted by the model, to adapt to searcher interests (Campbell and Van Rijsbergen, 1996).However, this aspect has not been investigated in the video retrieval domain. Two of the main state-of-the-art video retrieval systems—the Informedia system and the Físchlár system, only adopt a simple explicit feedback model for their query reformulation. The advantages of implicit feedback models are ignored completely. The main advantages and disadvantages have been introduced in section 1.3.3. For improving the performance of an interactive video retrieval system, it is a must to reduce or eliminate the negative effects of implicit feedback and explicit feedback and combine the positive effects of these two models.

1.4.2 Research Objectives

The aim of this thesis is to make an effective investigation of implicit feedback methods, implicit factors for interactive video retrieval and the approach to the combination of explicit and implicit features in an interactive video retrieval system. The assumption I make is that searchers will view information that relates to their needs; their interests can be inferred by monitoring what information they view. **The basic hypothesis proved is that users prefer an interactive video retrieval system with the combination of implicit and explicit features, but not an interactive video retrieval system with only explicit features.**

In this thesis, we present four novel implicit factors for the particular environment—video retrieval: selecting a result item which is a shot of a video; viewing a result item (view the key-frame, and text-based summary of a shot); playing a result item, duration of playing an result item, which will assist searchers in formulating query statements and making new search decisions on how to use these queries. Implicit feedback frameworks are created that use interaction with these factors and the traversal of paths between these factors as evidence to select terms for query modification and to make decisions on how to use the revised query.

The effectiveness of each of the proposed implicit feedback models and interest indicators were evaluated in the TREC Video Track framework. A simulation approach was used for investigating which implicit feedback model performs best among all of implicit feedback models proposed. Based on these results, the best performing model and most indicative implicit factor are chosen to be tested in a user experiment with human subjects.

Results of the simulated evaluation showed that the Binary Voting Model, which is a heuristic-based implicit feedback model (White *et al.* 2003) and the objective of which is to identify features for refining the query from the documents viewed by the user, performs really better than other variant models based on the Binary Voting Model and other alternative models (Ostensive Model, variant based on Ostensive Model). Therefore, in the subsequent user experiment, we use the Binary Voting Model to model user's action. Interface techniques are developed and tested that encourage interaction and aim to generate an increased quality and quantity of evidence for the implicit feedback methods devised. For two different Video RF systems (one is a system with only explicit features, the other has a combination of explicit and implicit features), we offer the same interface support, but the strategies of modelling the user's action are different. In the system with explicit features, we only modelled user's explicit action of marking relevance of video shots. In the system with a combination of explicit and implicit features, we modelled user's actions when using a video retrieval system, which are the four implicit factors proposed in this section, by using the Binary Voting Model. The results of user experiment proved my hypothesis that participants prefer the interactive system with combination of explicit and implicit features

In the rest of the thesis, firstly, related literature will be reviewed. Secondly, the basic framework of video retrieval systems I have developed will be introduced. Thirdly, a simulated approach to those implicit feedback models and factors will be presented, and the experiment results will be explained. Subsequently, the user experiments will be presented and related results will be analysed and a conclusion will be provided from this work.

Chapter 2

Interactive Video Retrieval Systems

The previous chapter provided an introduction to video retrieval. In this chapter, I will review basic components of an Information Retrieval Systems from five different perspectives—Retrieval Model, Interface support for IR systems, Relevance feedback, Query Categorization and Evaluation of an IR system. For each aspect, I will review the relevant literature followed by a review of corresponding applications. In the following, CMU, DCU and UNC will be used to refer to Carnegie Mellon University, Dublin City University and The University of North Carolina at Chapel Hill, respectively.

2.1 Retrieval Models

2.1.1 Vector Space Model

The origin of vector space model is derived from text retrieval. It is based on the assumption that, in some sense, the meaning of a document can be represented by a vector which represents words in the document. This makes it possible to compare documents with queries, which are represented by a vector, to determine how similar their content is. As in (Salton, 1975), the vector space model computes a measure of similarity by defining a vector that represents each document, and a vector that represent the query. Although this concept is derived from text retrieval, it is also appropriate to be used in multimedia information retrieval, including image retrieval, or video retrieval and so on.

The simplest way to construct a binary vector is to place a one in the corresponding vector component if the term appears, and a zero, if the term does not appear (Grossman, 1998). This scheme for the construction of a vector is too simple for more complicated document collection and ignores the importance of the words in the documents. Some words are more important than others and this is not reflected in binary vector scheme. The word, e.g. "chemistry", is more indicative than a general word-"element" in a given context, though both of these two words can be the elements of the vector. There are two main approaches of assigning term weights, one is to weigh terms manually by users, the other is to weigh terms automatically by IR system, typically "based on the frequency of a term as it occurs across the entire document collection" (Grossman *et al.*, 1998) . It is simply said that frequency-based weighting scheme is that a term that occurs infrequently should be weighed higher than a term that occurs frequently. One of the most popular weighting methods is TF-IDF (Rijsbergen, 1979). The formula of it goes as follow:

$$d_{ij} = tf_{ij} \times idf_j \qquad (2.1)$$

tf_{ij} is the number of occurrences of term t_j in document $d_{i;}$

idf_j is log(d/df_j) where d is the total number of documents[inverse document frequency],

d_{ij} is the weight of term t_j in document d_i.

d_{ij} is the number of times the term j appears in the document i.

Another variant (2.2) has been identified as a good performer (Salton, 1989):

$$w_i = \frac{(\log tf_{ij} + 1.0) * idf_j}{\sum_{i=1}^{n} [(\log tf_{ij} + 1.0) * idf_j]^2} \qquad (2.2)$$

INQUERY system (Salton *et al.*, 1989) adopted another useful weighting strategy. The weight of a term is computed using the INQUERY weighting formula, which uses Okapi's *tf* score (Robertson, Walker, and Jones, 1995) and INQUERY's normalized *idf* score:

$$w_i = \frac{tf_{ij}}{tf_{ij} + 0.5 + 1.5\dfrac{doclen}{avgdoclen}} \bullet \frac{\log(\dfrac{N+0.5}{docf})}{\log(N+1)} \qquad (2.3)$$

where tf_{ij} is the number of times the term occurs in the document, *docf* is the number of documents the term occurs in, *doclen* is the number of terms in the document, *avgdoclen* is the average number of terms per document in the collection, and *N* is the number of documents in the collection.

Similarity Measures

There are several different ways of comparing a query vector with a document vector. In all of these measures, the most common measure is the cosine similarity, idea of which is that the cosine of the angle between the query and document vectors is the quantitative value for the similarity between the query and the document. The formula is:

$$SC(Q, D_i) = \frac{\sum_{j=1}^{t} w_{qj} d_{ij}}{\sqrt{\sum_{j=1}^{t} d_{ij}^2 \sum_{j=1}^{t} w_{qj}^2}} \qquad (2.4)$$

Since the $\sqrt{\sum_{j=1}^{t} w_{qj}^2}$ appears in the computation for every document, the cosine coefficient should give the same relevance results as the traditional method- the inner product by the magnitude of the document vector. This method of computing similarity is general and can be suitable for computing similarity between two image feature vectors with a higher dimension.

2.1.2 Review the Application of Vector Space Model

Because of the generality of Vector Space Model, it can be used in many situations. The research groups from CMU, DCU and UNC used vector space to model their video retrieval system.

Informedia system of CMU (Hauptmann, 2003), runs manual search by exploiting multiple retrieval agents in the dimensions of color, texture, ASR (Automatic Speech Recognition), OCR(Optical Character Recognition), and some of the classifiers (such as anchor, PersonX (a classifier which is to filter video shots related to a single person)), which are represented by n-dimension vectors. Text-based baseline system used the OKAPI BM25 retrieval formula (Robertson, Walker, and Jones *et al.*, 1995).

Físchlár is a web-based video retrieval system, in which there are two essential aspects of the retrieval and weighting scheme for the system: the text search aspect and the image search aspect. Both image-based search engine and text-based engine use vector space to model the video shots space. Particularly, the image search engine measures similarity between an image-based query and all video shots in the video collection by computing image query dissimilarity. A Grouping Of adjacent Shots (GOS stand for Grouping Of adjacent Shots), consisted of five shots, is presented for improving the performance of search (Browne *et al.*,2001).

Open-Video project also applied image-based features and text-based features for video retrieval guided by TRECVID2003. Three different systems were evaluated. The transcript-only system allowed users to search the ASR transcripts of the video collection via a text box for search entry. The MySQL text search engine, which takes into account the number of words in a record, the number of unique words in that record, the total number of words in the collection, and the number of records that contain a particular word, was used fully. The search results were ranked based on the relevance score computed by MySQL, which uses a variant of the classic formula (Singhal, Buckley and Mitra, 1996), and adds on some calculations for "the normalization factor" for computing the weights of terms, and uses the product of term weight and the number of times the term appears in the query (MySQL *et al.*).

The features-only system allowed users to search the features provided from ten groups' results of the TREC VID 2003 features extraction task, results of which were aggregated by generating a "features score" on each feature for each shot; the score was the proportion of the runs that identified that feature in a particular shot. The 17 features were represented to users as semantically-related groups of items with checkboxes. The meanings of the features were provided in a training handout, and users were allowed to check as many features as they liked. The results from this system were ranked based on the average feature score for each shot, across all features included in the search. The third system provided both transcript and features searching, and required that users enter at least one term and check at least one feature. They received the instructions combined from the other two systems (Wildemuth, Yang, Hughes, Gruss, Geisler, and Marchionini, 2003).

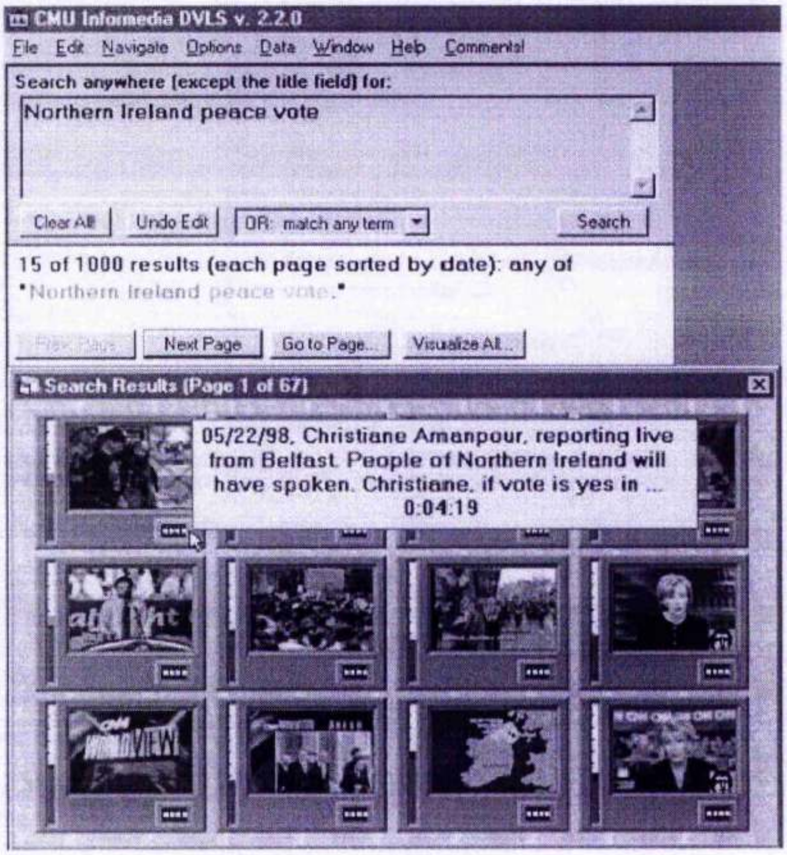
2.2 Interface Support

User interface is one of the most important components in a digital video system. It builds a bridge of communication between system users and system. Particularly in a video retrieval system, it transfers information needs of users to the search engine and presents the results of retrieval. In this section, I will report on some research interfaces to video retrieval systems.

2.2.1 Informedia system

As described above, the Informedia system is a digital video library with the explicit goal of enabling full-content search and retrieval for the full motion video and many modalities that video encapsulates. The aim of the Informedia interface was designed to provide users with quick access to relevant information in the digital video library. In order to help users decide which video they wanted to see the Informedia system presented the provision of alternative browsing capabilities--multimedia abstractions, which included headlines, thumbnails, filmstrips, and skims (Wactlar, Christe, Gong, Hauptmann, 1999).

Figure 2.1 Search and Result Panel of Informedia System


CMU Informedia DVLS v. 2.2.0
File Edit Navigate Options Data Window Help Comments!
Clear All
Undo Edit
OR: match any term
Search
Prev Page
Next Page
Go to Page...
Visualize All...
Search Results (Page 1 of 67)

Search anywhere (except the title field) for:

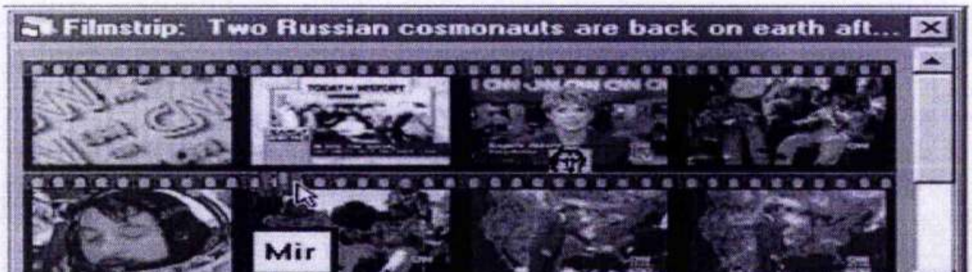
Northern Ireland peace vote

15 of 1000 results (each page sorted by date): any of "Northern Ireland peace vote."

05/22/98. Christiane Amanpour, reporting live from Belfast. People of Northern Ireland will have spoken. Christiane, if vote is yes in ... 0:04:19

Figure 2.1 shows the Informedia search interface following a query on the Northern Ireland peace treaty vote. The top of the search interface shows the terms of a text-based query by a text box in which multi-line area plain text will be displayed. The various operations between terms are implemented by a combo box. The display at the bottom shows thumbnail images for video segments returned as matches for the query. When the user positions the mouse arrow over a thumbnail, the interface pops up a headline for the segment (Wactlar, Christe, Gong, Hauptmann *et al.*, 1999).

Figure 2.2 Filmstrip and video playback windows for a result from the query for "Mir collision"



Filmstrip: Two Russian cosmonauts are back on earth aft...

Two Russian cosmonauts are back on earth after tro...

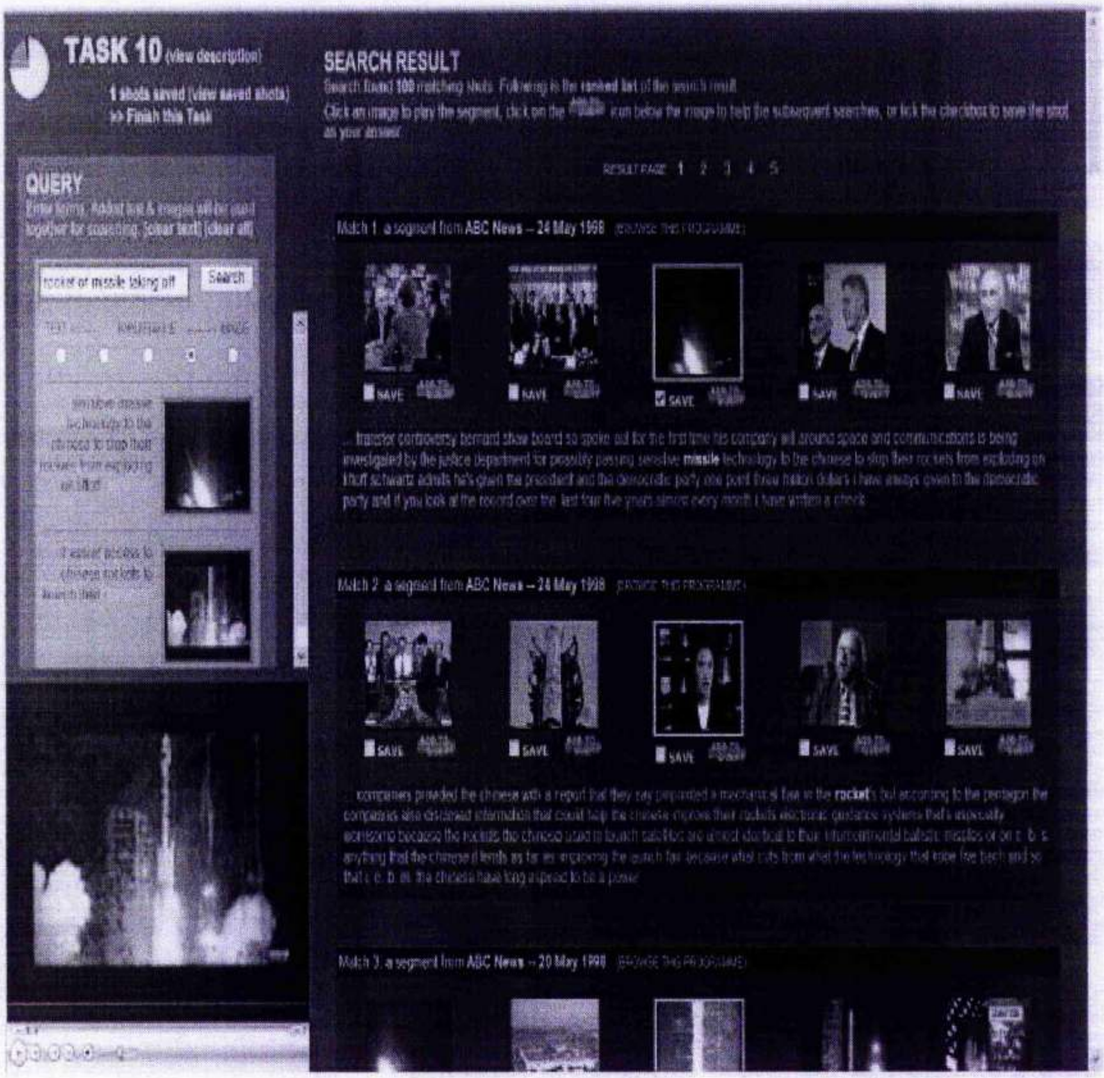
stay, they endured some of the most serious accidents in the history of the Mir space station including a fire

At the top of Figure 2.2, it is easy to show that the filmstrip is very useful to help identify key shots with bars color-coded to specific query words, in this case red for "Mir" and purple for "collision." because there is a segment's filmstrip which quickly shows the segment that contains more than a story matching the query "Mir collision," including an opening sequence and a weather report. Traditional media player is used to playback the currently selected video segment. Spoken transcripts text appears at the bottom of the video playback window. As the video plays, text scrolls are highlighted and spoken. The interface worked well with showing a single type of tightly synchronized metadata: spoken transcripts (Wactlar, Christe, Gong, Hauptmann *et al.*, 1999).

2.2.2 Físchlár

Físchlár has a web-based interface. Figure 2.3 show the main interface of Físchlár system. The search interface is positioned at the top-left of main interface.

Figure 2.3 Físchlár Main Interface



TASK 10 (view description)

1 shots saved (view saved shots)

>> Finish this Task

SEARCH RESULT

Search found 100 matching shots. Following is the ranked list of the search result.

Click an image to play the segment, click on the [illegible] icon below the image to help the subsequent searches, or tick the checkbox to save the shot as your answer

RESULT PAGE 1 2 3 4 5

QUERY

Enter terms. Added text & images will be used together for searching. [clear text] [clear all]

rocket or missile taking off

Match 1: a segment from ABC News – 24 May 1998 (BROWSE THIS PROGRAMME)

□ SAVE □ SAVE ☑ SAVE □ SAVE □ SAVE

... transfer controversy bernard shaw board so spoke out for the first time his company will around space and communications is being investigated by the justice department for possibly passing sensitive **missile** technology to the chinese to stop their rockets from exploding on liftoff schwartz admits he's given the president and the democratic party one point three million dollars i have always given to the democratic party and if you look at the record over the last four five years almost every month i have written a check

Match 2: a segment from ABC News – 24 May 1998 (BROWSE THIS PROGRAMME)

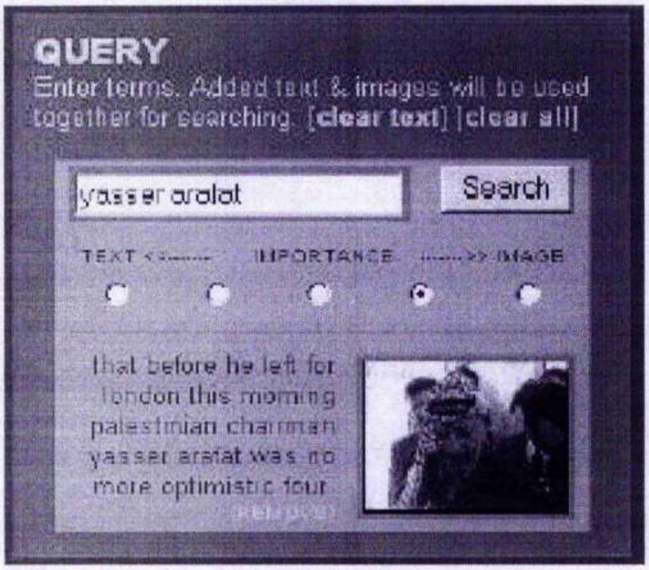
□ SAVE □ SAVE □ SAVE □ SAVE □ SAVE

companies provided the chinese with a report that they say pinpointed a mechanical flaw in the **rocket**'s but according to the pentagon the companies also disclosed information that could help the chinese improve their rockets electronic guidance systems that's especially worrisome because the rockets the chinese used to launch satellites are almost identical to their intercontinental ballistic missiles or i c b m s anything that the chinese learned [illegible] that i c b m the chinese have long aspired to be a power

Match 3: a segment from ABC News – 20 May 1998 (BROWSE THIS PROGRAMME)

Figure 2.4 shows a snapshot of the search interface. A text box is used to input terms of a text-based query. Five radio buttons are used to indicate which kind of feature is more indicative of relevance. Right side radio buttons of the middle radio button indicates that image features are more important, the left side radio buttons of middle radio button indicate that textual features have higher importance. The image, which is used to build a query by image example, is showed at the bottom of this interface. It is the key-frame of a video shot which is marked as relevant by user. Related text-based description of this video shot is showed on the left hand of the image (Browne *et al.*, 2001;Lee *et al.*, 2001).

Figure 2.4 Search Panel of Físchlár



QUERY

Enter terms. Added text & images will be used together for searching. [**clear text**] [**clear all**]

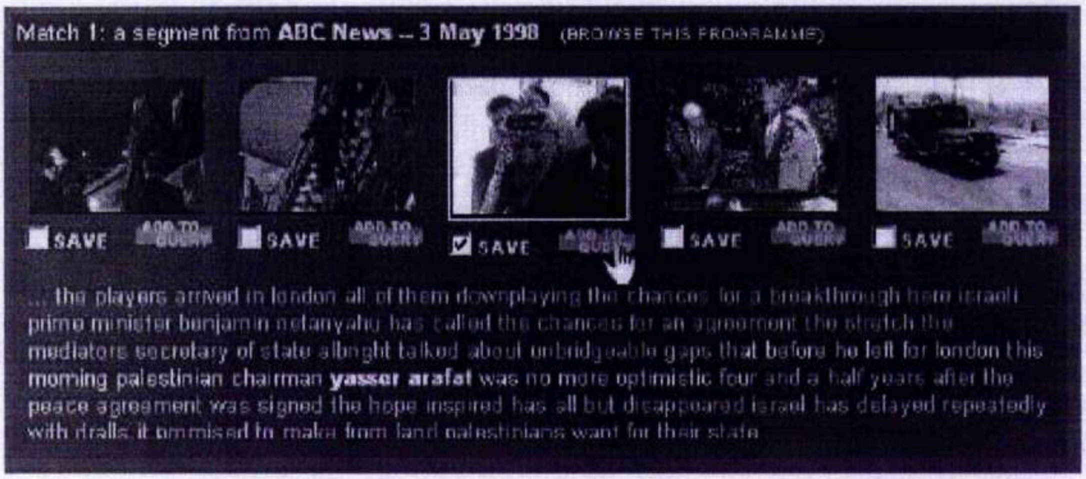
yasser arafat

TEXT <<------ IMPORTANCE ------>> IMAGE

○ ○ ○ ● ○

that before he left for london this morning palestinian chairman yasser arafat was no more optimistic four

Figure 2.5 Search Results



Match 1: a segment from **ABC News -- 3 May 1998** (BROWSE THIS PROGRAMME)

■ SAVE ■ SAVE ☑ SAVE ■ SAVE ■ SAVE

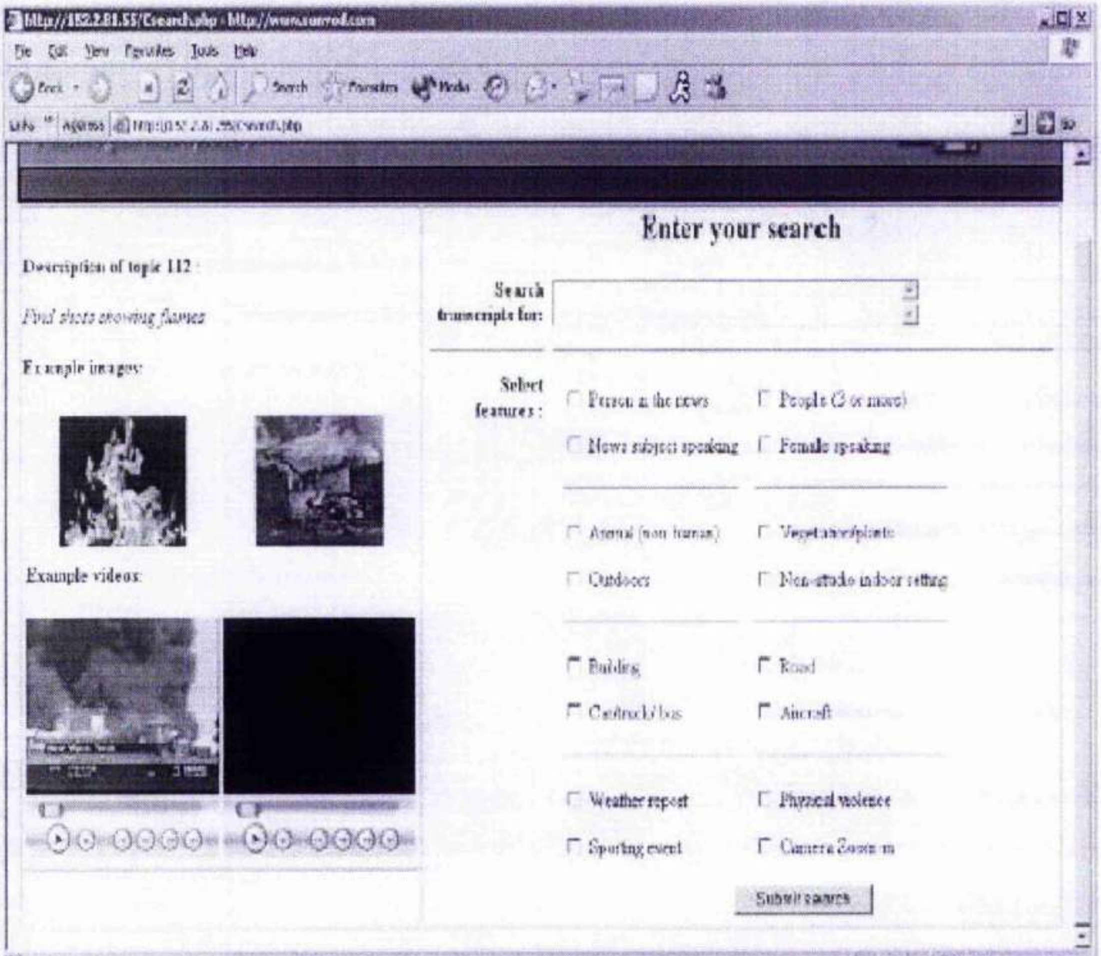
... the players arrived in london all of them downplaying the chances for a breakthrough here israeli prime minister benjamin netanyahu has called the chances for an agreement the stretch the mediators secretary of state albright talked about unbridgeable gaps that before he left for london this morning palestinian chairman **yasser arafat** was no more optimistic four and a half years after the peace agreement was signed the hope inspired has all but disappeared israel has delayed repeatedly with drafts it ommised to make from land palestinians want for their state

Figure 2.5 shows the interface of showing results of running a query. Thumbnails, which display key-frames of video shots, are showed for 5 video shots as well as Informedia System. The 'Add to Query' button below a key-frame is used to adds that shot content (text and image) into the search interface and subsequent search will use this shot along with the initial text term used. Standard Windows Media Player, which is located at the left-bottom of the main interface, is used for playback video shots (Browne *et al.*,2001; Lee *et al.*, 2001)

2.2.3 Open-Video Project for TRECVID2003

The system of Open-Video Project, which is used for TRECVID2003, is also a web-based video retrieval system. Figure 2.6 shows that the search interface is positioned at the right side. User is asked to input terms of a text-based query. 17 checkboxes are used to represent 17 features as semantically-related groups of items. The system also uses standard Microsoft Windows Media Player for playback video shots.

Figure 2.6 UNC System Main Interface



Enter your search

Description of topic 112 :

Find shots showing flames

Example images:

Example videos:

Search transcripts for:

Select features :

□ Person in the news □ People (3 or more)

□ News subject speaking □ Female speaking

□ Animal (non human) □ Vegetation/plants

□ Outdoors □ Non-studio indoor setting

□ Building □ Road

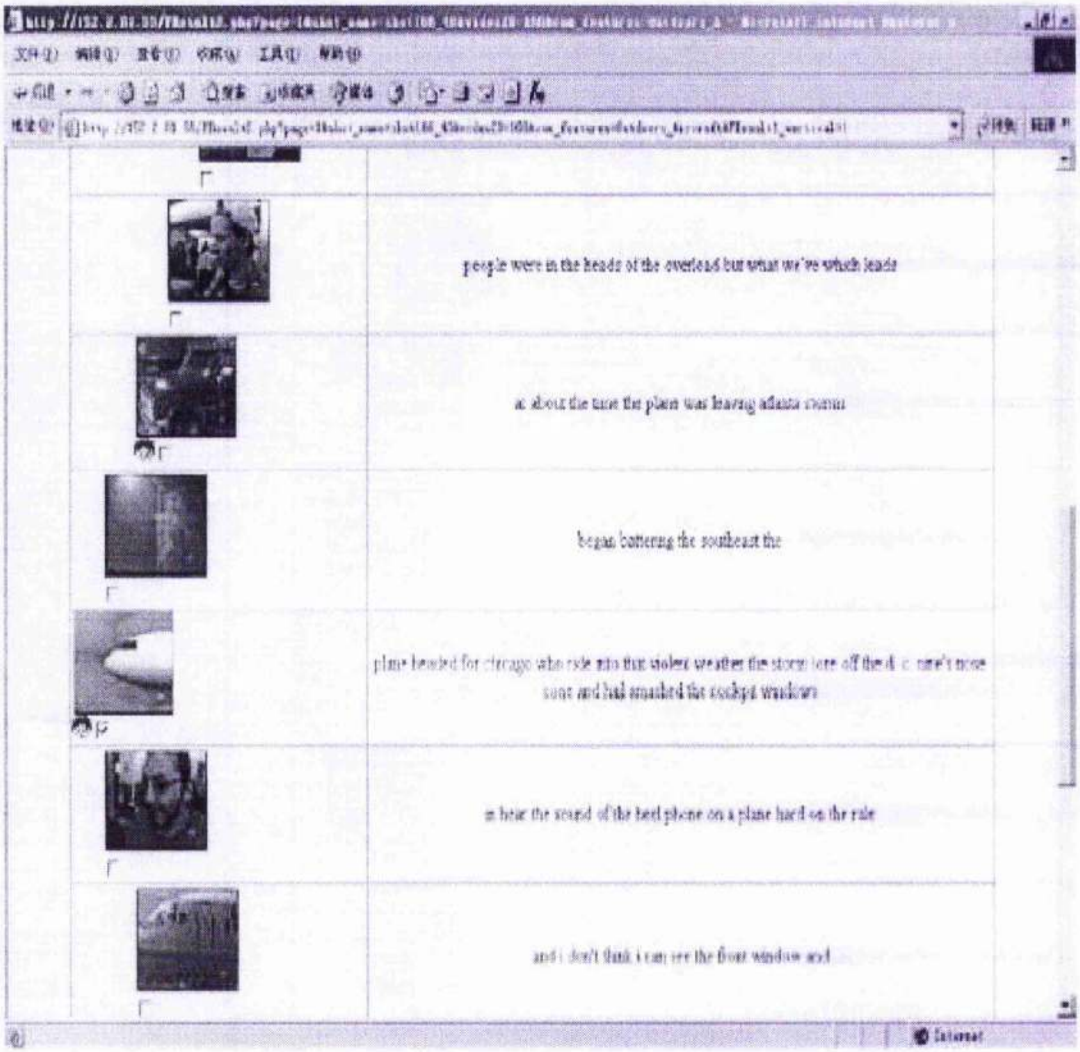
□ Car/truck/bus □ Aircraft

□ Weather report □ Physical violence

□ Sporting event □ Camera Zoom in

Figure 2.7 shows the interface of presenting search results (Wildemuth, Yang, Hughes, Gruss, Geisler, and Marchionini *et al.*, 2003) . The results were displayed, by default, in a horizontal view which includes a key-frame from each shot plus a few words from the transcript, selected in a window surrounding the search terms. After clicking on the key-frame in either of these basic views, the user will go to a before-and-after view, which shows video shots preceding and following the selected shot are represented in this view by their key-frames and full transcripts. The key-frame of the selected shot is aligned on the left side of the column, with the before and after shots indented slightly (Wildemuth, Yang, Hughes, Gruss, Geisler, and Marchionini *et al.*, 2003).

Figure 2.7 Search Results of UNC System



people were in the heads of the overload but what we've which leads

at about the time the plane was leaving atlanta storms

began battering the southeast the

plane headed for chicago who rode into that violent weather the storm tore off the d c nine's nose cone and had smashed the cockpit windows

m hear the sound of the bed phone on a plane hard on the rule

and i don't think i can see the front window and

2.2.4 Summary

In this section, I summarize the review of interface support for video retrieval systems from the following perspectives:

- Search interface
- Result interface
- Play interface

2.2.4.1 Search Interface

From these three systems, it is easy to see that the text box is presented to users for the input of search terms. All of these three systems use a text box for the input of search terms. In the Informedia system, a user can use the combo box to indicate the operations working on search terms. Físchlár system provides the interface support for a query-by-image-examples(use an image as a query) by user's relevance assessment. The system also allows users to indicate the relative importance of two kinds of features by five radio buttons. UNC's system also involves image samples into the retrieval process. 17 semantically-related feature groups were used for search by 17 checkboxes (Person in the news, People(3 or more), News subject speaking, Female speaking, Animal(non-human), Vegetation/plants, Outdoors, Non-studio indoor setting, Building, Road, Car/truck/bus, Aircraft, Weather report, Physical violence, Sporting event and Camera Zoom-in). A potential problem with such an approach is that selecting appropriate semantic groups is difficult for users.

2.2.4.2 Result Interface

All of these systems show the search results by using thumbnails of key-frames. In the Informedia system, a color bar is with each thumbnail which shows relevance of a video shot to the initial query. Headline will be showed when a user position mouse arrow over a thumbnail. The Informedia System also presents a new interface--Filmstrip for helping identify key shots. Key frames from a segment's shots can be presented in sequential order as filmstrips. The Físchlár system provides a checkbox for each video shot retrieved initially, which allows users to have an opportunity to explicitly mark the relevance of each video shot. Marked video shots are added to the search interface and used in the subsequent query. The basic result view of UNC's system looks like an annotated storyboard, and includes a key-frame from each shot plus a few words from the transcript, selected in a window surrounding the search terms. It presents a before-and-after view when a user click a key-frame in the basic view, the key-frame of the selected shot is aligned at the left side of the column, with the before and after shots indented slightly.

2.2.4.3 Play Interface

All of these three systems use standard Microsoft Windows Media Player to playback video shots. However, the Informedia highlights the transcript texts which are spoken while playing a video segment.

Through reviewing these three most effective in terms of performance video retrieval systems, it is easy to see that the basic interface elements—search interface, result interface and playback interface, are essential for a video retrieval system. The function of the search interface is to build a query (text terms, images, or combination of these two), the result interface is aimed at showing video shots retrieved by a query and at providing the support for relevance assessment, and the playback interface is used to provide the basic control operations, such as play, stop and pause playing video segments. All elements are absolutely necessary. However, forms of interface support are different.

2.3 Relevance Feedback

In the classic model, a query is devised and submitted by the searcher. Searchers are typically expected to describe the information they require via a set of query words submitted to the search system. This query is compared to each document in the collection, and a set of potentially relevant documents is returned. The query is a one-time static conception of the problem, based on the assumption that the information need remains constant for the entire search session. It is rare that searchers will retrieve the information they seek in response to their initial retrieval formulation (Rijsbergen, 1986). However, such problems can be resolved by iterative, interactive techniques. The initial query can be reformulated during each iteration either explicitly by the searcher or based on searcher interaction.

Relevance feedback (RF) is a technique that helps searchers improve the quality of their query statements and has been shown to be effective in non-interactive experimental environments and to a limited extent in Interactive IR (Beaulieu *et al.*, 1997). It is suggested that RF is an iterative process to improve a search system's representation of a static information need, which means the need after a number of iterations is assumed to be the same as at the beginning of the search (Bates *et al.*, 1989). The aim of RF is not to provide information that enables a change in the topic of the search (White, 2005).

Relevance feedback, originally developed for textual document retrieval (Rocchio, 1971), is a post query technique used to improve the effectiveness of information system, which uses positive and negative examples weighed from user to improve system performance. From the aspect of user's interaction, relevance feedback can be divided into two main types---one is the explicit feedback and the other is the implicit feedback. Though RF is originally developed for textual document retrieval, it can be used in the area of multimedia IR system for improvement of the performance of system.

A basic computing formula is formula (2.9) (Rocchio *et al.*, 1971) in this thesis.

$$Q_{i+1} = \alpha Q_i + \beta \sum_{rei}^{n} D_i / | D_i | - \gamma \sum_{nonrei}^{n} D_i / | D_i | \quad (2.9)$$

From this formula, we can see that the former query is the base of the new query. The last two parts of this formula shows that it considers positive and negative effect from examples marked by user during the process of query reformulation. α, β and γ are suitable constants, pre-set by system developer. Salton gave us suggestions about these three parameters, $\alpha=\beta$=0.75, γ= 0.25; or β=1, γ= 0; $\beta=\gamma$=0.25 (Salton and Buckley *et al.*, 1990).

These three different settings focus on different emphases. The first one focuses on more consideration of positive examples, less negative; the second does not consider the effect of negative examples; the last considers that positive and negative examples make the same effort to the query reformulation.

2.3.1 Explicit Feedback

Explicit feedback model asks the user to mark explicitly the relevance of documents in the results for improving the effectiveness of IR system. It is necessary that user makes assessments for the relevance of initial results. The main user interaction in this kind of model is to explicitly mark a document with various forms as relevant or non-relevant with, which is simpler than implicit feedback model. The interface for explicit feedback must provide the functions for user to check or mark which document is relevant to the query. It is stated that the interface support for explicit RF can often take the form of checkboxes next to each document at the interface, allowing searchers to mark documents as relevant, or a sliding scale that allows them to indicate the *extent* to which a document is relevant (Ruthven, Lalmas, and Van Rijsbergen , 2002b).

According to users' assessments of relevance of documents, the IR system reformulates the former query and re-searches by using the reformulated query (Salton and Buckley *et al.*, 1990). A number of studies have found that searchers show a desire for explicit relevance feedback features and, in particular, term suggestion features. Beaulieu and Walker 1992 evaluated an automatic query expansion (AQE) facility in the Okapi System and showed benefits of explicit relevance feedback. Koenemann 1996 investigated the use and effectiveness of an advanced information retrieval (IR) system (INQUERY), and suggested that the availability and use of relevance feedback increased retrieval effectiveness. He also suggested that the increased opportunity for user interaction with and control of relevance feedback made the interactions more efficient and usable while maintaining or increasing effectiveness by offering different level of interaction with a relevance feedback facility. Belkin (2000) suggested that explicit term suggestion is a better way to recommend system support for query reformulation.

However, Beaulieu *et al.*, 1997; Belkin *et al.*, 2001; Ruthven, 2001 indicated that the features of RF systems are not used in interactive searching; there appears to be an inconsistency between what searchers say they want and what they actually *use* when confronted with RF systems.

2.3.2 Implicit Feedback

As the previous sections have demonstrated, the problems with Explicit RF systems make effective alternatives appealing. Implicit feedback techniques unobtrusively infer information needs based on search behaviour, and can be used to personalize system responses and build models of system users. Implicit feedback techniques have been used to retrieve, filter and recommend different types of document (e.g., Web documents, email messages, newsgroup articles) from a variety of online sources. The primary advantage in using implicit techniques is that they remove the cost to the searcher of providing feedback (Nichols, 1997). Implicit measures are generally thought to be less accurate than explicit measures, but if implemented carefully can be effective substitutes for them (White *et al.*,2002b).

Categorization of Implicit feedback behaviours

Since implicit feedback is based on searcher behaviour there can be many possible sources for implicit evidence. (Nichols *et al.*, 1997; Oard and Kim, 2001; Claypool, Le, Waseda and Brown *et al.*, 2001; Kelly and Teevan, 2003) all provide conceptual classifications of potential behavioural sources of implicit feedback. Nichols (*et al.*, 1997) proposed the first classification of implicit feedback by categorising the actions that a searcher might be observed performing during information seeking and discusses the costs and benefits of using implicit ratings in information seeking, and categorises these ratings by the actions a searcher may perform.

Based on Nichols's work, Oard and Kim (Oard and Kim *et al.*, 2001) categorised observable feedback behaviours into four behaviour categories (Examine, Retain, Reference and Annotate), which refers to the underlying purpose of the observed behaviour, and also define minimum scope of these basic behaviours (Segment, Object and Class), which refers to the smallest possible scope of the item being acted upon. Table 2.1 shows the basic category of simple observable behaviours.

Table 2.1 Potentially observable behaviours

Minimum Scope

Behavior Category	Segment	Object	Class
Examine	View Listen	Select	
Retain	Print	Bookmark Save Delete Purchase	Subscribe
Reference	Copy-and-paste Quote	Forward Reply Link Cite	
Annotate	Mark up	Rate Publish	Organize

According to the above figure, 'Examine' is where a searcher studies a document, and examples of such behaviour are view (e.g., reading time), listen and select. 'Retain' is where a searcher saves a document for later use and examples include bookmark, save and print. Further examples of keeping behaviours on the Web, where information is retained for later re-use, can be found in (Jones, 2001). 'Reference' behaviours involve users linking all or part of a document to another document and examples include reply, link and cite. 'Annotate' are those behaviours that the searcher engages in to intentionally add personal value to an information object, such as marking-up, rating and organising documents.

Kelly and Teevan (*et al.*, 2003) provide us an extension of the category of observable behaviours. The new 'Create' category describes the behaviours typically associated with the creation of original information.

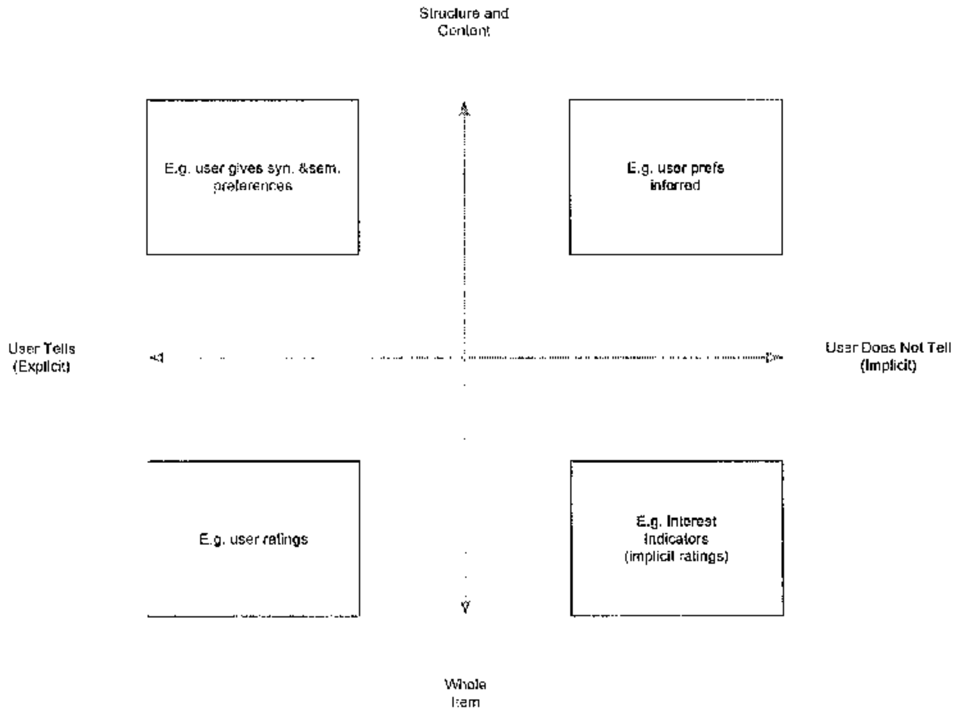
All of five categories are sufficient to classify most search behaviour, though those only represent a subset of the possible behaviours that searchers may perform. In all of those categorises, 'Reference', 'Annotate' and 'Create' categories all require control over the content of documents and the structure of document spaces, but the 'Examine' and 'Retain' categories are appropriate to categorise the behaviour of online searchers because searchers rarely have this control.

Table 2.2 Classification of implicit behaviours (Oard and Kim *et al.*, 2001) **with the additions added by Diane Kelly and Jaime Teevan.**

Behavior Category	*Minimum Scope*		
	Segment	*Object*	*Class*
Examine	*View* *Listen* *Scroll* *Find* *Query*	*Select*	*Browse*
Retain	*Print*	*Bookmark* *Save* *Delete* *Purchase* *Email*	*Subscribe*
Reference	*Copy-and-paste* *Quote*	*Forward* *Reply* *Link* *Cite*	
Annotate	*Mark up*	*Rate* *Publish*	*Organize*
Create	*Type* *Edit*	*Author*	

In Claypool, Le, Waseda and Brown *et al.*, 2001, authors address a categorization of different interest indicator categories, including explicit and implicit based on their customized browser, which can record the online behaviour, used as implicit measures of interest.

Figure 2.8 Categorizing interest indicators


Structure and Content
E.g. user gives syn. &sem. preferences
E.g. user prefs inferred
User Tells (Explicit)
User Does Not Tell (Implicit)
E.g. user ratings
E.g. Interest indicators (implicit ratings)
Whole Item

The categorization of Claypool, Le, Waseda and Brown *et al.* 2001 is a two-dimension representation of all interest indicators. The horizontal of axis of it represents the degree of explicit or implicit of the interest indicators, and the vertical axis represents the source of indication-- the structure or content of the item or from whole item. The area from the bottom middle to bottom right of the Figure represents the implicit interest indicators. They also provide another categorization for it: Explicit Interest indicators, Marking Interest Indicators, Manipulating Interest indicators, Navigation Interest Indicators, External Interest Indicators, Repetition Interest Indicators, and Negative Interest

Indicators. All of these interest indicators are context sensitive, the dependency of which is user's task or goal. The performance directly relies on different combinations of all these interest indicators.

Review of the application of user's implicit evidence

Claypool, Le, Waseda and Brown *et al.* 2001 examined the actions: mouse click, scrolling, and time on browsing. Different actions were measured in different ways. "Mouse click and scrolling were measured both as the number of mouse click and as total time spent". Scrolling can be also measured both at the keyboard and with the mouse. Experimental subjects were asked to browse documents in an unstructured way. The time spent on a page, mouse clicks and scrolling were all recorded automatically by the customised browser that subjects used. Subjects were asked to explicitly rate each page before leaving it and the ratings were used to evaluate the implicit measures. The researchers found a strong positive correlation between time and scrolling behaviours and the explicit ratings assigned. However, since subjects were not engaged in a search task and just asked to browse a set of interesting documents, the applicability of the findings to information seeking scenarios is uncertain.

Morita and Shinoda (1994) proposed observations of reading time as the implicit interest indicators. They obtained a strong positive correlation between reading time and explicit feedback provided by the eight users. These users were required to read all articles posted to the newsgroups of which they were members and to explicitly rate their interest in the articles for six weeks. There are very low correlations between the length of the article and reading time, the readability of an article and reading time and the size of the user's news queue and reading time. Several reading time thresholds for identifying interesting documents were examined and applied to experiments which resulted in the finding of the most effective threshold--20 seconds, resulting in 30% of interesting articles being identified at 70% precision (Morita and Shinoda *et al.*,1994).

Golovchinsky, Price and Schilit (1999) used the text generated by a user as the implicit evidence of user interests. They constructed full text queries based on users' annotated passages of documents and compared their IR system. The system is based on the construction of full text queries users' annotated passages of document, which can provide the system with a more refined, user-specific unit, with which to perform relevance feedback and help in establishing a context. It is better than using just a list of terms, to standard relevance feedback techniques. They concluded that the performance of their system was better than the standard one (Golovchinsky, Price and Schilit *et al.*, 1999).

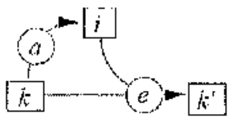
Budzik and Hammond (1999) proposed an interest indicator—URL to the user based on what the user was typing. The result of their experiments proves that the implicit indicator they suggested is really useful and performs better (Budzik and Hammond *et al.*, 1999). Kleinberg (1999) improved the performance of his system by the large-scale use of the analysis of Web link.

2.3.3 The Ostensive Feedback Model

The Ostensive Feedback Model is derived from the theory of development of information needs (Campbell and Van Rijsbergen *et al.*, 1996). It is a model of learning that is used to continue updating knowledge state.

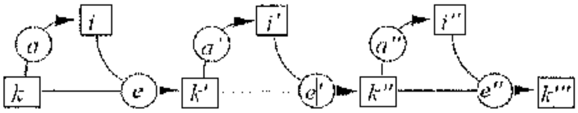
The following Figure 2.9 and 2.10 illustrate the basic components and procedure of development of information needs of human beings based on Ostensive Model.

Figure 2.9: The updating of a knowledge state through the selection of, and subsequent exposure to, information.



Where k denotes knowledge state of user; *i* denotes information object; *a* with a circle denotes action (selection); *e* with a circle denotes exposure process based on the effect of learning information object *i*; k' denotes the next knowledge state of user through the process of exposure.

Figure 2.10 The iterative updating of a knowledge state.



Where K, K', K'', K''' are different knowledge state, e is the process of interpreting i with respect to, or within, a context k. This figure clearly illustrates that the complete process of development of information needs is an iterative updating process of a knowledge state, which is a good and simple reflection of the process of development of information needs in real life. The number of times of iterative updating is increased or decreased according to particular conditions.

The ostensive model can also be defined as a model that recognises the changing uncertainty inherent in a user's cognition of his information need. But what is Ostensive definition? The ostensive definition in the area of Philosophy is that "the explanation of a word by presenting, pointing at, or otherwise indicating one or more objects to which it applies". In this definition, the term 'word' is taken as a denotation of the abstract notion 'relevance to an information need' (Campbell and Van Rijsbergen *et al.*, 1996).

The three underlying elements of ostension are defined in Campbell and Van Rijsbergen *et al.* 1996):

- Pure ostension: equates to simple observed evidence,

- Identification: refers to the recognition of identity of the concepts being defined by the individual acts of pure ostension.

- Induction: the process of combining the evidence.

The centre of this model is focused on the collecting and combination of ostensive evidence. Authors take uncertainty as the indication of ostensive evidence and define several basic types of uncertainty profiles, which describe the relationship between the degree of uncertainty and age: a decreasing profile of uncertainty (Figure 2.11), a flat profile of uncertainty (Figure 2.12), an increasing profile of uncertainty (Figure 2.13), a decelerating profile, and an accelerating increase in uncertainty with age (Figure 2.14).

Figure 2.11 A decreasing profile of uncertainty

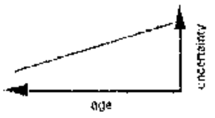


Figure 2.12 A flat profile of uncertainty

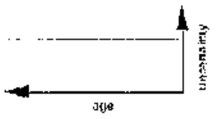


Figure 2.13 An increasing profile of uncertainty

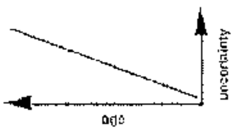
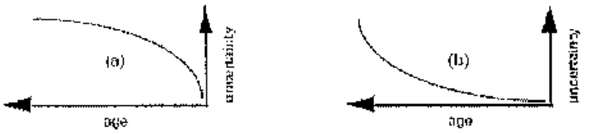


Figure 2.14 A decelerating profile, and an accelerating increase in uncertainty with age.



The increasing profiles of evidence uncertainty indicate that the most recent evidence has the lowest attached uncertainty and therefore will have the most influence on the ostensive definition. Here, all ostensive evidence plays a part in the ostensive definition; nevertheless, the most recent will play the greatest. This means that the ostensive definition will follow recent trends in the ostensive evidence, but will always have a component of the historical evidence. The decreasing profiles indicate that old evidence is more indicative of the current knowledge state than more recent evidence. This means that the early evidence has the most influence on the ostensive definition, and that subsequently observed evidence becomes of less and less importance.

2.3.4 Review relevance feedback models of the three systems (Informedia, Físchlár ,Open-Video system)

Informedia system

In TRECVIDEO 2002 search task, Negative Pseudo-Relevance Feedback (NPRF) was applied to Informedia system. NPRF, choose the unlabeled data farthest from positive data as the negative sample, which is also a common strategy used in some earlier work of positive-based learning or self-learning based on an underlying assumption that positive data are more likely to be in the boundary of the data set (Hauptmann *et al.*,2003). It was also proved to be effective at providing a more adaptive similarity measure. For TRECVID2003 a modified version of NPRF score based on the idea of original NPRF algorithm, which combined Maximal Marginal Relevance (MMR) criterion proposed by (Carbonell, Geng, and Goldstein, 1997), which takes both "relevance", "irrelevance" and "novelty" into account.

Físchlár:

In the Físchlár system, it is allowed to explicitly mark the relevance of the video shots. The system allows a user to add any shot's content into the subsequent queries if that shot was felt to contain relevant visual content or relevant text content. In this way, the relevance feedback mechanism could be used to expand a query using any video shots encountered by the user during interaction. Image query dissimilarity was used to process image-based queries (Browne *et al.*,2001; Lee *et al.*, 2001).

Open-Video System

The current Open-Video System does not apply any relevance feedback models to improve the performance of video retrieval (Wildemuth, Yang, Hughes, Gruss, Geisler, and Marchionini *et al.*, 2003; Marchionini and Geisler *et al.*, 2002).

2.4 Query Categorization

Classification of search tasks has been widely investigated in the community of information retrieval and query answering. Li and Roth (2002) presents a machine learning approach to question classification, which guided a hierarchical classifier by a layered semantic hierarchy of answer types, and eventually classifies questions into fine grained classes. VideoQA system explores the use of question answering (QA) techniques to support personalized news video retrieval by adopting a hierarchical classification approach to categorize free-form factual queries, which classified the questions into 8 main question classes (or answer targets) (Yang, Chaisorn, Zhao, Neo and Chua, 2003). They are Human, Location, Organization, Time, Number, Object, Description and General. The last group, General, is used to group questions that cannot be categorized into other classes. Five types of machine learning approaches, which include Nearest Neighbors (NN), Naïve Bayes (NB), Decision Tree (DT), Sparse Network of Winnows (SNoW), and Support Vector Machines are experimented for automatic question classification task (Zhang and Lee, 2003). Kang, and Kim (2003) classifies the user queries into three categories, that is, the topic relevance task, the homepage finding task and the service finding task using various statistics from query words. Different linear weights of text information and hyperlink information will be assigned based on the query categories to improve the web document retrieval. The similar idea can be naturally extended to the context of video retrieval. Rong, Yang, and Hauptmann (2004) proposed using query-class dependent weights within a hierarchical mixture-of-expert framework to combine multiple retrieval results. Firstly, they classify each search tasks defined by TRECVID2003 into one of the four pre-defined categories: Named person (P-query) queries for finding a named person, possibly with certain actions, Named object (E-query) queries for a specific object with a unique name, which distinguishes this object from other objects of the same type, General object (O-query) queries for a certain type of objects, Scene (S-query) queries depicting a scene with multiple types of objects in certain spatial relationships (Rong, Yang, and Hauptmann *et al.*, 2004).

From the research of the application of query categorization describe above, it is proved that the query categorization is really helpful for information retrieval and query answering. But the three main video retrieval systems (Informedia, Físchlár ,Open-Video system) did not investigate the most appropriate features (low-level features, semantic features, and textual features) based on the categories of queries for video retrieval. In my study, I categorize search tasks to quantify different features for video retrieval.

2.5 Evidence Combination: Linear Combination, Dempster-Shafer theory, and Voting

A simple linear combination of scores was originally proposed by Porkaew, Chakrabarti and Mehrotra (1999). Fagin, Kumar, and Sivakumar (2003) proposed an aggregation method purely based on ranks—(*Voting Approach*). In the *Voting Approach (VA)*, each query representative is treated as a voter producing its own individual ordering of candidates (images). The *median rank aggregation* method was mainly used to compute the combined list. They rank the database elements based on similarity to the query by using a small number of independent "voters". These rankings are then combined by a highly efficient aggregation algorithm. Our methodology leads both to techniques for computing approximate nearest neighbors and to a conceptually rich alternative to nearest neighbors. The algorithm MEDRANK was proven to be very efficient and database friendly by their two sets of experiments.

The Dempster-Shafer (DS) Theory of Evidence Combination is also a powerful framework for the combination of results from various information sources, and has been extensively studied for IR purposes (Jose,1998).

2.6 Evaluation of IR Systems

I will review the current trends in the evaluation of IR Systems. The simulated methodology will be proposed in the following section.

2.6.1 Evaluation of IR Systems

It is very important to evaluate IR systems because of the social and economic factors (Rijsbergen *et al.*, 1979). Traditionally, IR systems can be evaluated from the following six perspectives (Cleverdon, Mills and Keen, 1966):

1) The coverage of the collection, that is, the extent to which the system includes relevant matter;

2) the time lag, that is, the average interval between the time the search request is made and the time an answer is given;

3) the form of presentation of the output;

4) the effort involved on the part of the user in obtaining answers to his search requests;

5) the recall of the system, that is, the proportion of relevant material actually retrieved in answer to a search request;

6) the precision of the system, that is, the proportion of retrieved material that is actually relevant.

However, for evaluating interactive IR systems, relevance is one of the most significant thins which should be considered. According to Rijsbergen *et al.*, 1979, relevance is a *subjective* notion. Different users may differ about the relevance or non-relevance of particular documents to given questions. However, the difference is not large enough to invalidate experiments which have been made with document collections for which test questions with corresponding relevance assessments are available.

Cleverdon (1960) used collections of documents, queries and pre-determined relevance assessments to determine the performance of indexing techniques and algorithms of the IR system. The Text Retrieval Conference (TREC) creates test collections and recruits assessors to assign relevance assessments to documents based on the approach used in Cranfield (Harman, 1993). The measurement of precision/recall was considered as a relevance-based measure of effectiveness that typifies a system-driven approach to developing and testing IR systems in controlled environments (Swanson,1986; Spärck-Jones, 1981). Buckley stated that RF algorithms were tested using similar methods and a very simple model of searcher interaction based on the simulated assessment of the top-ranked documents (Buckley *et al.*, 1994). But Belkin and Vickery argued that those approaches are restrictive and do not model searcher interaction fully and make assumptions that places limits on the cognitive and behavioural features of the environment in which IR systems operate (Belkin and Vickery, 1985). It is simple to say that searchers interact or the processes involved in the interaction is neglected and not evaluated.

The new approaches, which are the models combining system-centred evaluation models with user-centred evaluation models, not only requires a comprehensive understanding of the nature of information systems but also need to completely know about the characteristics of information needs and relevance assessments by individuals. New approaches (task-oriented) take into account actual or simulated information seeking environments. In (Hersh, Elliot, Hickam, Wolf, and Molnar, 1995), authors described an approach to evaluate the usefulness of information retrieval systems. A measure, which is to ask medical students to answer questions from a shot answer test, was used in their evaluation of two information retrieval systems.

Egan (1989) reported a formative evaluation of a hypertext system called SuperBook, which is hypertext browsing system designed to improve the usability of accessing electronic documents. A number of questions, including open ended questions were designed to emulate various kinds of usage of such a system, and students with a background in statistics were used as subjects to test their system.

Giorgio Brajnik, Stefano Mizzaro, and Carlo Tasso described the evaluation of IR interface (FIRE) based on different tasks/topic combinations. 45 computer science undergraduate students were invited as subjects and asked to use a different system to resolve each problem in a related-sample, within-subjects design.

Lancaster (1996) proposes an approach for the evaluation of interactive knowledge-based systems. They compare the effect of indexing produced by the use of MedIndEx (an expert system with indexing produced through an automated indexing management system) with National Library of Medicine's (NLM) system for indexing. They asked 60 inexperienced indexers and 20 experienced indexers to index the same 30 medical documents.

Borlund and Ingwersen describe the ideas of assumptions underlying the development of a new method for the evaluation of interactive IR systems, which takes into account the dynamic nature of the information needs which are assumed to develop over time for the same user, and is designed to involve real users with simulated work task situation (Borlund and Ingwersen, 1997).

These studies raise the following common concerns:

i) Meaningful evaluation of the whole range of a search's interaction with systems.

ii) The observation of the behaviour of 'real' users engaged in the evaluation.

iii) Performance criteria, not just relevance-based effectiveness

iv) Acquisition and analysis of data, qualitative, that may be used to measure the performance of systems, not like traditional measures.

2.6.2 Simulation-Centric Evaluation Methodology

Simulation-based methods have been used for the test of query modification techniques (Harman, 1998; Magennis, 1998; Ruthven, 1998). Therefore, the simulation based methods can also be regarded as another feasible methodology for evaluating the effectiveness of RF technology, because Simulated-centric methods are less time consuming and less costly than experiments with human subjects, allow the comparison of IR techniques in different retrieval scenarios, and maintain control over environmental and situational variables. Tague-Sutcliffe and Nelson (1981) proved that a modified algorithm for the simulation of user relevance judgments, which integrated the physical as well as the logical and semantic elements of these systems, was validated in the bibliographic retrieval systems. Mostafa (2003)focused on the dynamic nature and the variability of user-interests and their impact on the modeling process by developing a simulation based information filtering environment called SIMSFITER to overcome some of the barriers associated with conducting studies on user-oriented factors that can impact interests. White proposed six different representatives and simulated user's search actions working on those representatives when using a web search engine (White, 2004).

Simulation-based methods have also been used among other things. For testing the usability of websites, Chi (2003) developed InfoScent™ Bloodhound Simulator which is a prototype service of automated usability tools based on the simulated-based strategy. It automatically analyzes the information cues on a Web site to produce a usability report. The algorithm of Information Scent Absorption Rate is used to measure the navigability of a site by computing the probability of users reaching the desired destinations on the site. Chi (2001) simulated Web searchers' action--the hyperlink clicks enable researchers to better understand the usage of the Web, designers to better design their Websites, and end-users to seek information more efficiently .

2.6.3 The Evaluation Methodologies of three video retrieval systems (Informedia, Físchlár, Open-Video system)

Informedia System

MAP (Mean Average Precision) measurement is used for manual search tasks in TRECVID2003. The interactive video retrieval evaluation of TRECVID2003 is used to evaluate the interface of the Informedia System (Hauptmann *et al.*, 2003). CMU conducted formal empirical studies to measure the effectiveness of particular multimedia abstractions and a number of evaluations on the system, including contextual inquiry, heuristic evaluation, cognitive walk-through, and think-aloud protocols. Usage data were tracked primarily by automatically logging mouse and keyboard input actions, supplemented with user interviews (Wactlar, Christe, Gong, Hauptmann *et al.*, 1999).

Físchlár

MAP is used for evaluating the performance and effectiveness of the system. In the user experiments, two variations of the Físchlár system were introduced in order to prove the hypothesis that the system with image search and relevance feedback mechanism outperforms the text-only system without the support of image search and relevance feedback. The basic elements of a user experiment were included. The training tutorial, the aim of which is to make subjects know how to use the two systems, was introduced to sixteen subjects. As with the official topics, suitable example images from the collection were also provided (Browne *et al.*,2001; Lee *et al.*, 2001).

Open-Video system

The method--user experiments was adopted as its main evaluation method. Thirty-six subjects were mainly from among students, faculty and staff at UNC. Posting flyers in several buildings on campus, as well as email announcements within the School of Information and Library Science were used to get all of subjects for the experiments. A research assistant was responsible for monitoring each session. A within-subjects research design was used, and all subjects were asked to use the system ready to be evaluated (Wildemuth, Yang, Hughes, Gruss, Geisler, and Marchionini *et al.*, 2003) .

Summary

All of these three video retrieval systems applied user experiment as one of evaluation methods for their systems. The traditional measuring method for evaluating the basic effectiveness of an IR system—Precision/Recall was also used. None of them did adopt the Simulation-Centric Evaluation Methodology for evaluations. Specially, the Físchlár system does not apply simulation centric strategy to evaluate the usefulness and effectiveness of relevance feedback and image search. However, Simulation-Centric Evaluation Methodology has been proved to be effective for the evaluation of usefulness and performance of relevance feedback mechanism in (White *et al.*, 2005; Harman *et al.*, 1998;Magennis *et al.*, 1998;Ruthven *et al.*, 1998).

Chapter 3

Query Categorization

In Chapter 2, I reviewed some approaches on query categorization. In this chapter, I present my approach to query categorization. Firstly, I describe an approach to study the relationship between query categorization and feature selection. Secondly, I propose the experimental results based on my approach and other approaches. A comparative study is provided.

3.1 Approaches to query categorization

In this section, Query categorization is approached from the following perspectives:

- Categorization of Search Tasks
- Role of Image features in Video Retrieval
- Role of text-based features
- Relationship between categories and image-low-level features and textual features

3.1.1 Categorization of Search Tasks

I used the video collection provided by TRECVID2003 and the associated queries (set 100-124) with rich text descriptions and categorized these search topics into the following four categories: person category, object category, scene category and event category. Person category contains the queries for finding a named person, possibly with certain actions (e.g., "find video shots of Yasser Arafat"). Object category contains the queries for a certain type of objects and refers to a general category of objects (e.g., "Find shots of the Sphinx"). Scene category queries are the queries which depict a scene with multiple types of objects in certain spatial relationships (e.g., "Find shots with aerial views containing both one or more buildings and one or more roads"). Event category queries are the queries for an event happening (e.g., "find shots of an airplane taking off"). There are five queries in the Person category, eight queries in the Scene category, six queries in the Object category, and six queries in the Event category.

Our definition of query classes is slightly different from the definition of query classes defined by Yan (Rong, Yang, and Hauptmann *et al.*, 2004).. The definitions of the categories of Person, Object and Scene are same as Yan's approach. With the exception that I combined the Named Object and General Object categories into one Object category. In addition, I define a category of Event which describes an action case. My approach to categorization of queries is based on the assumption that the forms of all queries are based on text and image features. For event category, an underlying action is essential. I felt that, such a category is significantly different from other categories and is needed in measuring the retrieval effectiveness.

The objective of this experiment is to measure the usefulness of various features for retrieval. Hence, we classified the query accordingly. Each query class defined by our approach is different from others and will reveal the usefulness of various features for retrieval. According to these two different categorizations, we can make a conclusion about the role of video features (low-level features and high-level features).

3.1.2 Role of Image Features in Video Retrieval

It is very important for a video retrieval system to select proper features for the computation of similarity. Low-level features such as color, texture, shape and so on, are the most essential features for representing a video. According to conclusions made by the user studies in UNC (Wildemuth, Yang, Hughes, Gruss, Geisler, and Marchionini *et al.*, 2003) and DCU (Browne *et al.*,2001), the semantic features (e.g. the feature extracted from ASR (Automatic Speech Recognition) results and Close-Caption) are effective in greatly improving the effectiveness of video retrieval systems. And the image features play effectively no role in video retrieval. These experiments are based on user study, however, no proper bench-marking study is performed.

In order to look into this problem and quantify the effect of image features in video retrieval, we designed the experiments based on my approach and Yan's approach (Rong, Yang, and Hauptmann *et al.*, 2004). The results and corresponding analysis will be presented in section 3.3.

3.1.3 Role of text-based features

The power of textual features is shown in conventional information retrieval systems. Compared to the low-level image features, they are high level in nature. That is, it has some semantic association. The main high-level source currently in use by video retrieval systems, which are not restricted to a specific genre, are speech transcripts, and these can be generated automatically from spoken audio or from closed caption information. A number of genre-specific features like object detection can also be extracted for domain-specific video content like television news, sports and cartoons. The text-based features are the most important features. Recently, more and more studies found that the textual features may be the most effective features for multimedia retrieval because of the semantic association (Wildemuth, Yang, Hughes, Gruss, Geisler, and Marchionini *et al.*, 2003; Browne *et al.*,2001).

3.1.4 Relationship between Query Categories and features--low-level features and textual features

Yan argued that each query class favored a specific set of features (Rong, Yang, and Hauptmann *et al.*, 2004). Hence, query-class dependent weights were used in a hierarchical mixture-of-expert framework to combine multiple retrieval results. The aim of his experiments was to develop a retrieval framework that uses class dependent weights for combing results from various features. His experimental results demonstrated that the performance with query-class dependent weights can be learned from the development data efficiently and can be generalized to the unseen queries easily (Rong, Yang, and Hauptmann et al., 2004). It also showed that the performance with query-class dependent weights can considerably surpass that with the query independent weights. However, his experiments did not show the performance difference of various features in respect to query classes. That is, no bench marking study is performed to find the effectiveness of various features with respect to the query categories. Our aim is to study this aspect.

For example, the name of a person is most critical to a search task of finding a person. Face presence, size, position information and face recognition based on low-level features may be also critical to such a kind of search task but of little value to other query classes. For the search topics of finding a person and finding a specific object, the transcript is particularly important since such queries are more likely to have perfect match in transcript. Since the specific object may have particular characteristics in various aspects (distribution of colour, the consistent direction, specific granularity, shape property and so on), the visual features, such as colour, texture, shape, and edge, may be significantly useful to improve the performance of such specific video retrieval. Therefore the idea of query class specific retrieval is generally applicable. Query classification, which captures query characteristics, is really helpful for the appropriate selection of features, which is critical for better performance. Our main objective is to conduct a comparative study measuring the effectiveness of various features and their combination.

3.2 Experimental methodology

In this section, I introduce the experimental methodology I used for the study of query categorization from two aspects:

- Query Categorization

- Strategy of weights for image features and text-based features

3.2.1 Query Categorization

The search topics used in this study are defined by TRECVID2003 (TRECVID, 2003) with identification number from 0100--0124.

Table 3.1 Query Categorization based on my approach

Person	Scene	Object	Event
	0100		
		0105	
	0101		0102
0103		0106	
	0108		0104
0114		0109	
	0112		0107
0118		0116	
	0113		0110
0119		0121	
	0115		0111
0123		0122	
	0117		0120
	0124		

Table 3.2 Query Categorization based on Yan's approach (Rong, Yang, and Hauptmann et al., 2004)

Person	Scene	Specific Object	General Object
			0104
	0100		
			0105
0103	0101	0108	
			0107
0114	0102	0124	
			0109
0118	0110	0106	
			0112
0119	0111	0116	
			0113
0123	0115	0120	
			0121
	0117		
			0122

3.2.2 Weights for image features and text-based features

In order to find a better combination of weights for compound query based on image features and text-based, we present five different combinations. The sum of the weight of image features and the weight of text-based features is 1.0. We did those experiments on full collection provided by TRECVID 2003 with no query classification and query also with classification (TRECVID *et al.*, 2003).

1. .3 and .7

2. .4 and .6

3. .5 and .5

4. .6 and .4

5. .7 and .3

For avoiding the unilateral results and conclusions, we also repeated experiments based on Yan's classification of queries by using the same setting of experiments mentioned above. We used global color histogram and the texture feature based on the Co-occurrence algorithm (Stricker and Orengo, 1995; Sonka, Hlavac and Boyle, 1998).

3.2.3 Evidence Combination

For combining the retrieval scores based on different kinds of features (text and image), I used two combination methods one is a weighted linear combination method, the other is based on the Dempster's evidence combination theory.

Linear combination

I use the linear combination method to combine two scores based on the two image features in the system. The weight for each image feature is 0.5 that means we think the importance of the two image features is same.

$$S_{img} = S_{color} \times 0.5 + S_{texture} \times 0.5 \qquad (3.1)$$

Where S_{img} is the final similarity score based on two image features (global histogram and Cooccurrence), S_{color} is the similarity score based on global histogram vector only, $S_{texture}$ is the similarity score based on Cooccurrence vector only.

$$S_{img} = S_{color} \times \mu + S_{texture} \times (1 - \mu) \ (3.2)$$

Where S_{image} is the final similarity score based on two image features (global histogram and Cooccurrence), S_{text} is the similarity score based on the textual feature, μ is the weight for color feature. It can take values .3, .4, .5, .6 and .7. In my study, an assumption is made that the importance of the two image features is same

Dempster-Shafer Theory of Evidence Combination

Two kinds of queries (term based query and image query) representing each feature are issued to the system, returning two result lists with different scores based on the respective similarity measure for each feature. A means to combine the results to obtain one single ranked list is the *Dempster-Shafer Theory of Evidence Combination*. The Dempster-Shafer mechanism has been widely used in the context of IR to combine information from multiple sources (Urban and Jose *et al.*, 2004). The advantage of Dempster's combination rule is that it integrates degrees of uncertainty or trust values for different sources (Urban and Jose *et al.*, 2004). For two features Dempster-Shafer's formula is given by:

$$m(\{d_i\}) = m_1(\{d_i\}) \times m_2(\{d_i\}) + m_1(\Theta) \times m_2(\{d_i\}) + m_2(\Theta) \ (3.3)$$

$$m\ (\Theta) = m_1(\Theta) \times m_2(\Theta) \ (3.4)$$

Where $m_k(\{d_i\})$ (for k = 1; 2) can be interpreted as the probability that document d_i is relevant with respect to source k. The two sources in our case correspond to the similarity values computed from the text and image feature respectively. Θ denotes evidence (also referred to as un-trust coefficients): $m_k(\Theta) = 1 - strength_k$ (3.5) where $strength_k$ the trust in a source of evidence (Urban and Jose *et al.*, 2004).

3.3 Experimental Results and Analysis

In this section, I presented the experimental results based on both my approach and Yan's approach to query classification from two aspects:

- Role of Image Features and Textual Features. In this experiment, the hypothesis is that textual features play more role than image features in video retrieval.

- Relationship between query categories and features. In this experiment, the hypothesis is that each category a different specific combination of features would be optimal.

3.3.1 Role of Image Features and Textual Features

Figure 3.1 shows the performance of the five different combinations of weights for image features and text-based features without regarding the query classification. In the following figures, I used the 0.3:0.7 (0.3 refers to the weight for image features and 0.7 refers to the weight for the text-based features) to refer to the combination of weights for image features and text-based features, and I used the linear combination approaches.

From this figure, we can see that the combination of 0.3 and 0.7 (image features and text-based features) has the best performance among these five combinations. With the decrease of the weight for text-based features, the performance decreases gradually.

Figure 3.1 Mean Average Precision/Recall

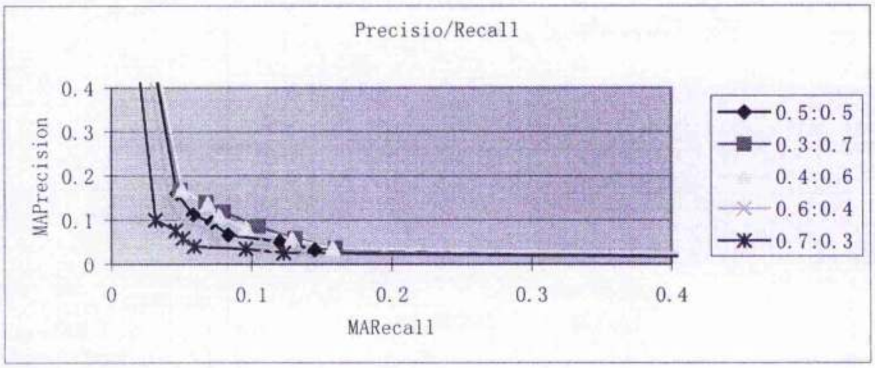


Table 3.3 Precision at Nth Document

Nth Document Combination of features (image : text)	10	20	30	50	100	200
0.3 : 0.7	**0.168**	**0.14**	**0.119**	**0.0856**	**0.0588**	**0.036**
0.4 : 0.6	0.168	0.136	0.109	0.0808	0.0572	0.0358
0.5 : 0.5	0.16	0.112	0.0947	0.0664	0.0508	0.0324
0.6 : 0.4	0.14	0.094	0.0827	0.0568	0.0416	0.0292
0.7 : 0.3	0.1	0.076	0.0587	0.04	0.0344	0.0256

Table 3.4 Recall at Nth Document

Nth Document Combination of features (image : text)	10	20	30	50	100	200
0.3 : 0.7	0.0491	0.0678	**0.08**	**0.105**	**0.131**	**0.16**
0.4 : 0.6	**0.0503**	**0.069**	0.077	0.0961	0.129	0.158
0.5 : 0.5	0.0467	0.0584	0.0715	0.0833	0.12	0.145
0.6 : 0.4	0.0414	0.0516	0.0683	0.0738	0.111	0.137
0.7 : 0.3	0.0315	0.0462	0.0512	0.0589	0.0959	0.123

3.3.2 Relationship between query categories and features

Figures 3.2-3.5 present the precision and recall figures according to the categorization we proposed. In the following figures, we use the terms img, and keywords to refer the use of image features, the use of text-based features respectively. The term 'linear' and 'dempster' to refer to the combination of both text-based feature and image features based on linear and D-S combination approaches respectively.

Figures 3.2-3.5 show that, basically a video IR system, using the text-based feature clearly outperforms the systems using image features only or both image features and the text-based feature. For the categories of Scene and Event, the performance of the system using text-based features is still higher than that of the other three systems based on image-only, linear combination, and D-S combination. However, for these categories combination of text and image features show closer performance to text features. The two approaches combining image features and text-based features based on different combination schemes have the same performance.

These four figures show that the text-based feature is the best to be used in a video retrieval system. However, it also demonstrated the use of image features for Event or Scene query. The results also show that the performance of the system using image features only is worse than that of the systems using compound features or using the text-based feature. There is no obvious difference when using linear combination method or the method based on Dempster-Shafer evidence combination. The results adequately prove the relationship between features and search topics we propose in this study. To summarize, our conclusions are that text features are superior for video retrieval. It makes sense to use the combination of features for the scene or event categories.

Figure 3.2 the Precision/ Recall curve of the category—Person

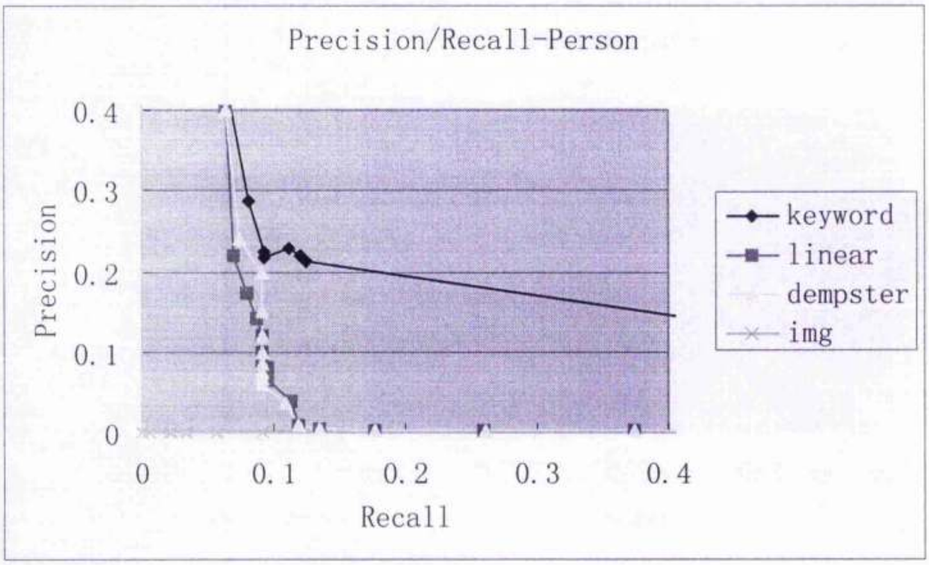


Figure 3.3 the Precision/ Recall curve of the category—Scene

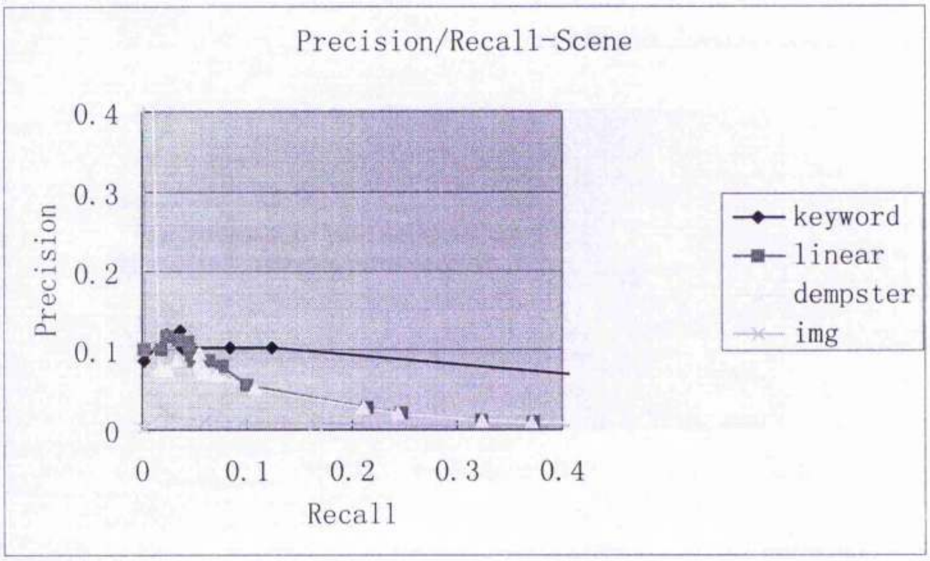


Figure 3.4 the Precision/ Recall curve of the category--Object

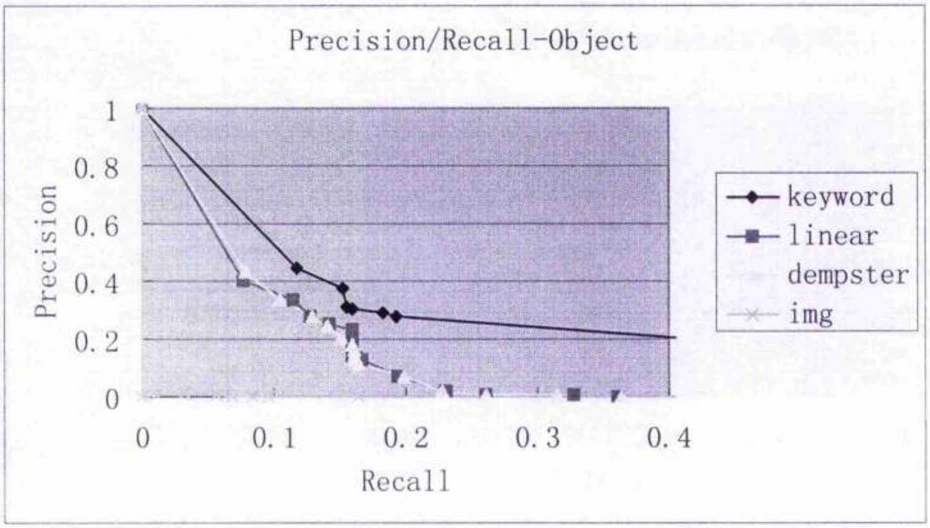
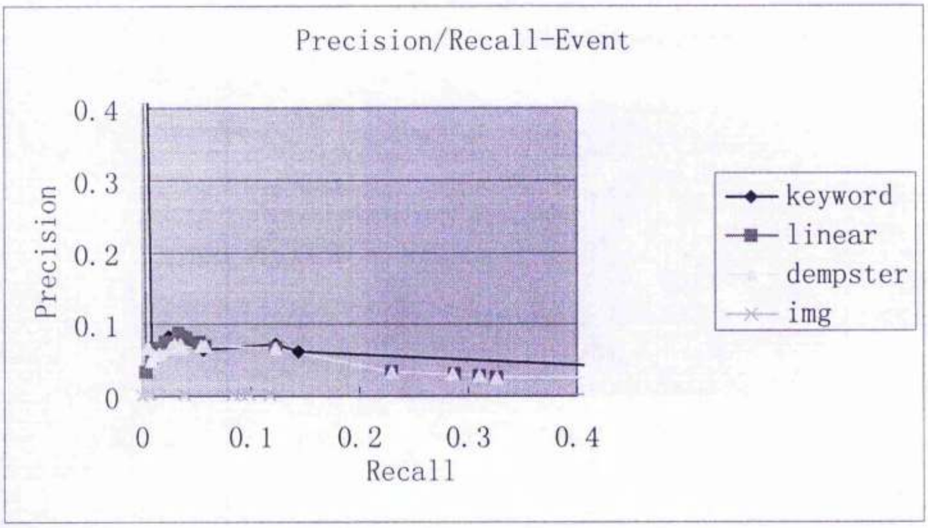


Figure 3.5 the Precision/ Recall curve of the category--Event



Results based on Yan's approach to query classification

Figures 3.6-3.9 present the precision and recall figures according to the categorization Rong, Yang, and Hauptmann (*et al.*, 2004) proposed. Since there is no difference between the linear combination method and the method based on Dempster-Shafer evidence combination, I use only the linear combination method in the following experiment.

Figures 3.6-3.9 show that, for a video IR system, use of the text-based feature can really improve the performance in comparison to systems using image features only. The Video IR system, using both image features and the text-based feature outperforms the system using image features only or text-based feature only for the category **Specific Object**.

For the Person, the system only based on text-based feature outperforms other two systems. For Scene and General Object queries, the system only based on text-based feature has the similar performance as the system based on both image features and text-based features. For the categories of Scene and General Object queries, the performance of the system using text-based is still higher than that of the other two systems. However, for these categories use of text and image features show closer performance to text features.

These four figures also show that the text-based feature is the best to be used in a video retrieval system. However, it also demonstrated the use of image features for three kinds of queries defined by Yan (Rong, Yang, and Hauptmann et al., 2004). The results also show that the performance of the system using image features only is worse than that of the systems using compound features or only using the text-based feature. It makes sense to use the combination of features for the Scene, and General Object or Specific Object categories, and it is more effective to only use text-based features for the Person Category.

Figure 3.6 Precision/ Recall curve of the category—Person

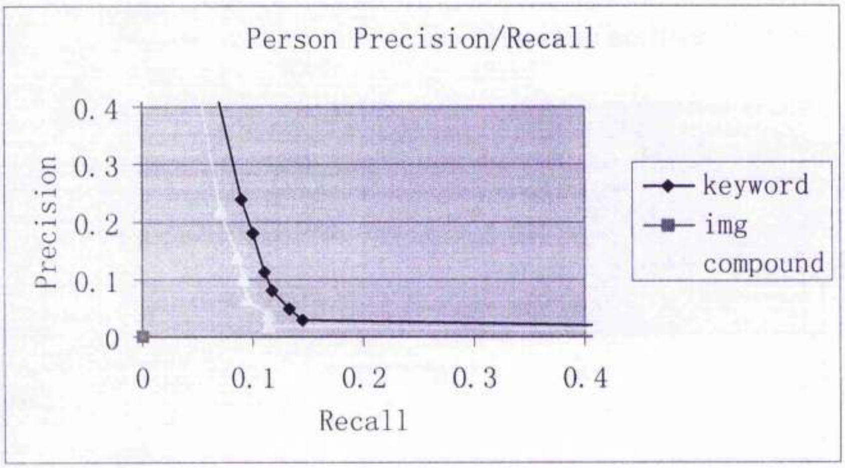


Figure 3.7 Precision/ Recall curve of the category—Scene

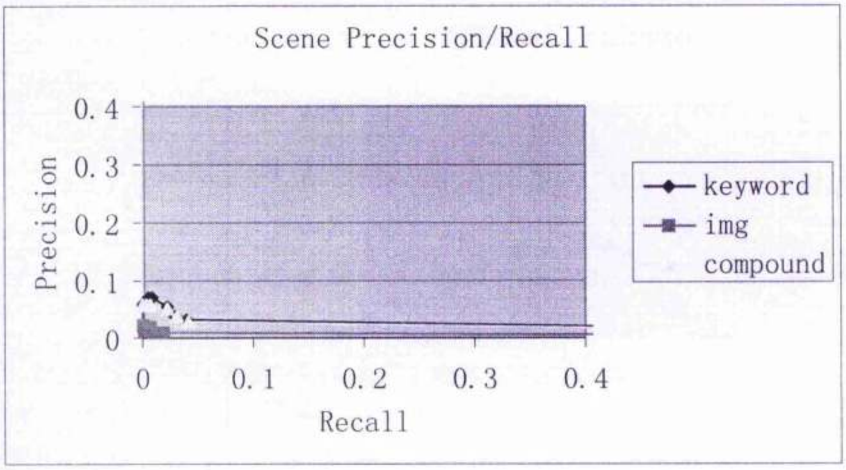


Figure 3.8 Precision/ Recall curve of the category—General Object

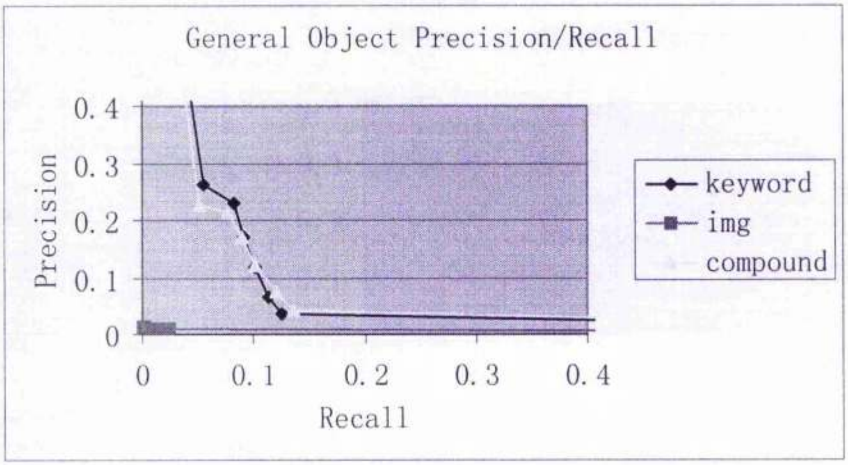
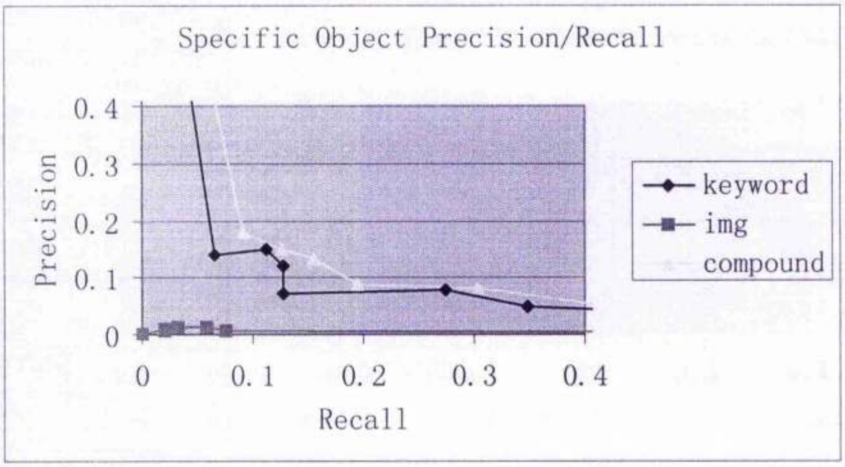


Figure 3.9 Precision/ Recall curve of the category—Specific Object



Summary

We have done experiments to quantify the effectiveness of various features for video retrieval. We categorized queries into various classes and experimented. In addition, we used a query categorization as proposed by (Rong, Yang, and Hauptmann et al., 2004).

From these experiments, we make the following conclusions. The most important feature for video retrieval is text features. This conclusion adheres to the general view in the field and also to the results of the user studies.

From the experiments using query categorization, we conclude that the image features are useful for the retrieval of video used in conjunction with textual features. This is true for the categories of scene and event.

We also experimented with two methods for combination of evidence. Our experimental results show that both of them perform in more or less the same fashion.

Chapter 4

JIVRSystem: A Prototype Video Retrieval System

In this chapter, we introduce an interactive video retrieval system called **JIVRSystem**. Designed and built by myself.This system is used to experiment in the later chapters. The architecture of the system and the component are discussed.

4.1 Video Collection

For development and also for the experimental purpose, the standard collection from TRECVID2003 was used. This collection includes 120 hours (241 30-minute programs) of ABC World News Tonight and CNN Headline News recorded by the Linguistic Data Consortium from late January through June 1998. The size of the files on the hardrive is a little over 100 gigabytes, which makes the evaluation based on TRECVID realistic collection, fair and practical. According to the guidelines of TRECVID2003 (TRECVID *et al.*,2003), the whole collection is divided into two parts: one is used to develop and tune the video retrieval system, the other is used to test the performance of video retrieval systems. In the JIVRSystem, the test part of the whole collection was used.

The videos in the collection have the following associated textual data:

- The output file (*.as1) of an automatic speech recognition system

- A closed-captions-based transcript

The transcript will be available in two forms:

- simple tokens (*.tkn) with no other information for the development and test data;

- tokens grouped into stories (*.src_sgm) with story start times and type for the development collection

- shot boundary information files(*.xml)

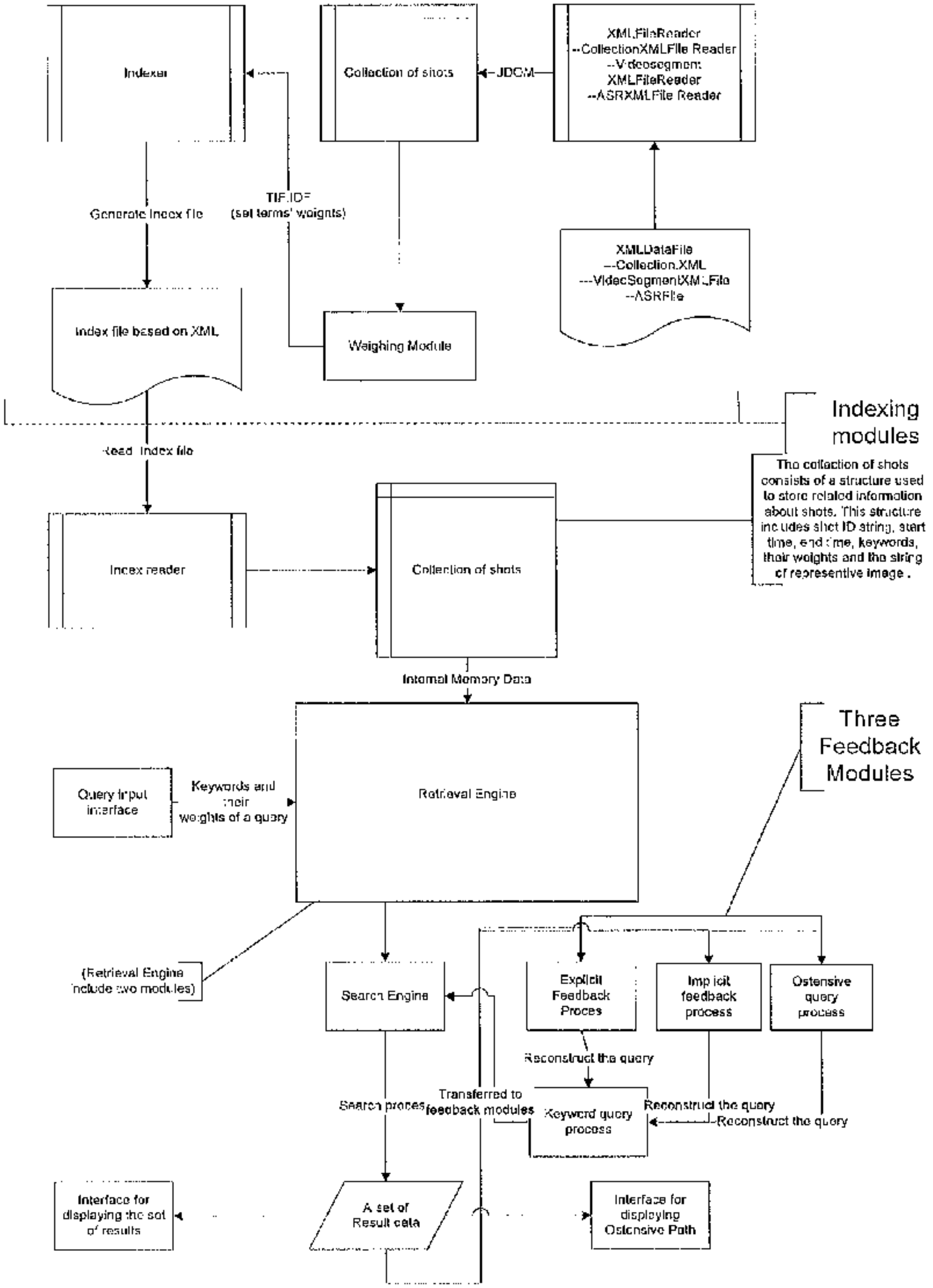
- a list of the files in the collection(collection.xml)

- a data set of the key-frames that are described in shot boundary information files.

4.2 System Architecture

The main components of interactive video retrieval systems are showed in following figure:

Figure 4.1 System Structure



This collection has an XML-based structure with its internal video description complying with MPEG-7 standard. Figure 4.1 shows the overall architecture of the basic system. The underlying descriptions about video documents and related shot boundary descriptions are based on XML. The ASR transcripts and close-caption texts are from LIMSI (LIMSI) results which are not based on XML. We developed a tool for indexing; all the shots described in related shot boundary descriptions are indexed. Our basic retrieval unit is a video shot. Two image features--Global Color histogram, Co-coocurrence texture (Sonka, Hlavac and Boyle et al., 1998), were extracted. The method of extracting text-based feature is described in the next section.

The subject submits a text-based query via a search panel the system provided. This panel processes it and sends it to the search engine. The search engine sends back the retrieved results that is a table of shots which show some information, such as shot name, shot duration, the key frame (represent image) of a shot, and extra-text-based descriptions. In the basic system, the interactive procedures are the most important modules. Our system consists of three main modules: Indexing modules, Retrieval Engine, Feedback Modules.

4.2.1 Indexing Modules

The systems use two distinct features: *text* and *visual*. The text feature is extracted from shot-based ASR Results generated by LIMSI (LIMSI). Visual features are extracted from image sample provided by TRECVID 2003. These features are stored based on XML format.

The Extraction of Textual feature:

The textual feature is extracted from ASR results. Firstly, the index module of JIVRSystem indexes the ASR texts into a set of words, and removes the stopwords. If a word is contained in the stop word list, the word cannot be a keyword (term), otherwise it could be a keyword and stored in the index.

The weight of a term is computed using the INQUERY weighting formula, which uses Okapi's *tf* score (Robertson, Walker, and Jones *et al.*, 1995) and Inquery's normalized *idf* score:

$$w_i = \frac{tf_{ij}}{tf_{ij} + 0.5 + 1.5\frac{doclen}{avgdoclen}} \bullet \frac{\log(\frac{N+0.5}{docf})}{\log(N+1)} \qquad (4.1)$$

where tf_{ij} is the number of times the term occurs in the document, *docf* is the number of documents the term occurs in, *doclen* is the number of terms in the document, *avgdoclen* is the average number of terms per document in the collection, and *N* is the number of documents in the collection.

The extraction of image features:

2 low-level features were extracted: global colour histogram, which uses 64-dimension vector to represent that image, the texture feature based on Cooccurrence algorithm which uses 20-dimension vector to characterize that image (Sonka, Hlavac and Boyle *et al.*, 1998).

4.2.2 Similarity Measure

The role of Retrieval Engine is to retrieve shots from video collection according to the text-based query. The similarity between a pair of documents or between th equerry and a document is measured by one over the cosine of the angle between the corresponding vectors, which is widely used in the vector-space model (Salton, 1989).

The formula is showed in (4.2):

$$SC(Q,D_i)=\frac{\sum_{j=1}^{t} w_{qj}d_{ij}}{\sqrt{\sum_{j=1}^{t} d_{ij}^2 \sum_{j=1}^{t} w_{qj}^2}} \qquad (4.2)$$

For filtering the appropriate number of shots, we define a threshold for the score, which can be set dynamically by the user interface and saved in the file system.

4.3 User Interface

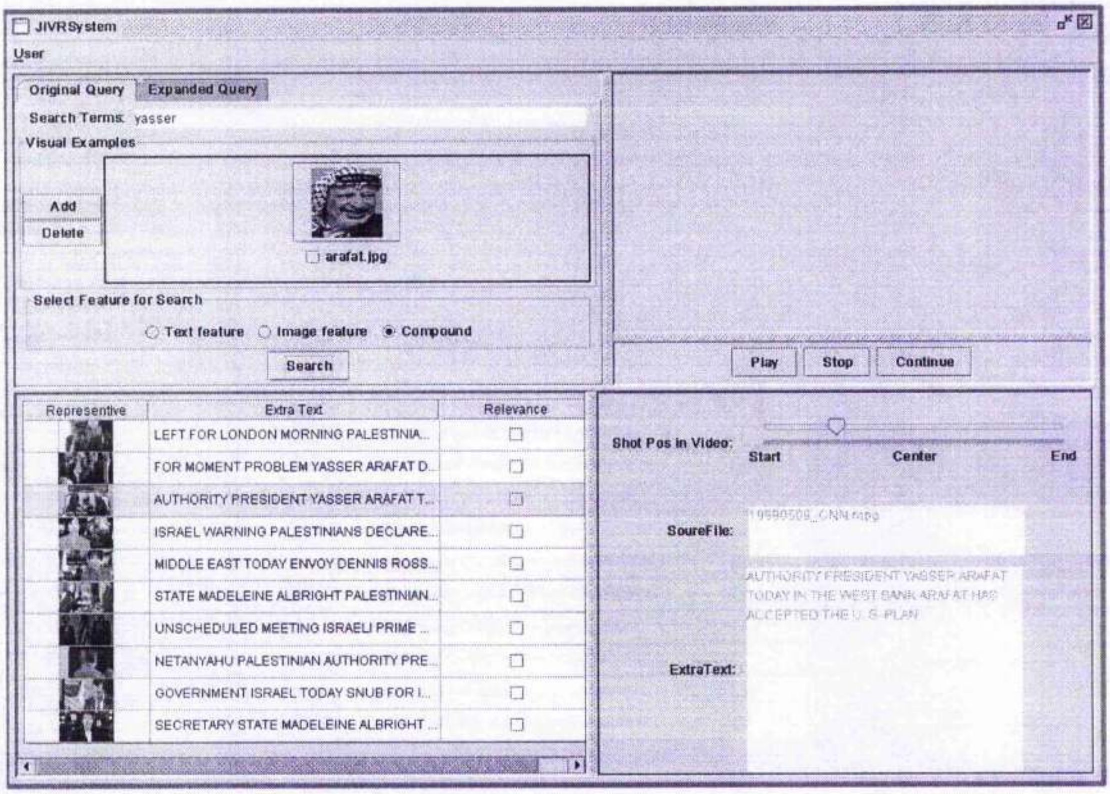
The interface of JIVRSystem can be divided into four parts: Search, PlayBack, Results Display and the interface for showing related Information of video shots. Figure 4.2 shows the main interface of the JIVRSystem.

In the following sections, I will introduce each of them in turn.

4.3.1 Search Interface

There are two panels for search interface: Original Query Panel, and Expanded Query Panel. They are tabbed panels and not visible simultaneously.

Figure 4.2 Main Interface of the JIVRSystem



Search Terms: yasser

Visual Examples

□ arafat.jpg

Select Feature for Search

○ Text feature ○ Image feature ● Compound

Representive	Extra Text	Relevance
	LEFT FOR LONDON MORNING PALESTINIA...	□
	FOR MOMENT PROBLEM YASSER ARAFAT D...	□
	AUTHORITY PRESIDENT YASSER ARAFAT T...	□
	ISRAEL WARNING PALESTINIANS DECLARE...	□
	MIDDLE EAST TODAY ENVOY DENNIS ROSS...	□
	STATE MADELEINE ALBRIGHT PALESTINIAN...	□
	UNSCHEDULED MEETING ISRAELI PRIME ...	□
	NETANYAHU PALESTINIAN AUTHORITY PRE...	□
	GOVERNMENT ISRAEL TODAY SNUB FOR I...	□
	SECRETARY STATE MADELEINE ALBRIGHT ...	□

Shot Pos in Video:

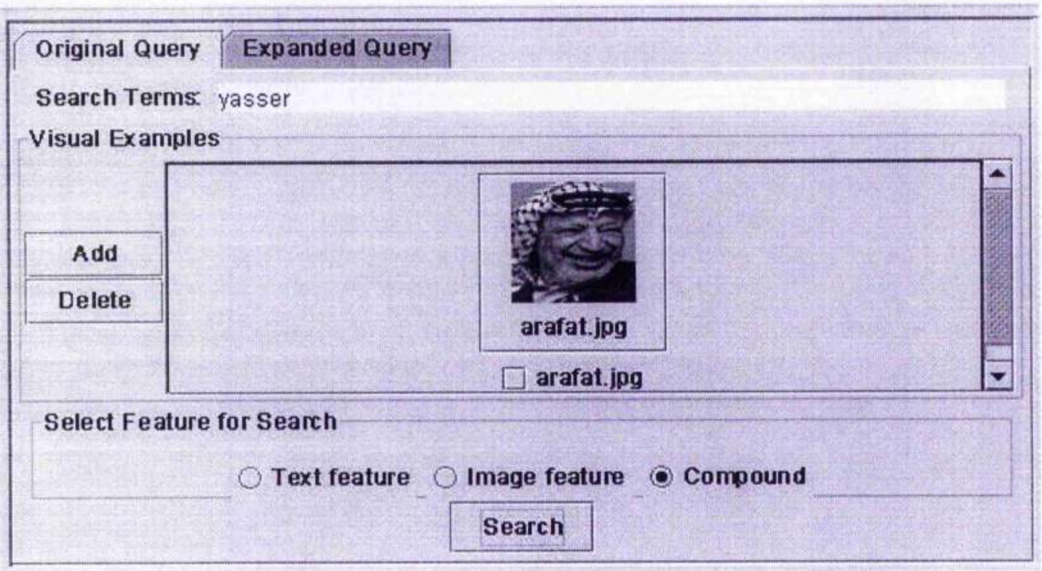
SoureFile: 19980509_CNN.mpg

ExtraText: AUTHORITY PRESIDENT YASSER ARAFAT TODAY IN THE WEST BANK ARAFAT HAS ACCEPTED THE U. S. PLAN

4.3.1.1 Original Query Panel

Original Query Panel is for constructing and showing an original query (Figure 4.3):

Figure 4.3 Search Panel



Search Terms: yasser

Visual Examples

□ arafat.jpg

Select Feature for Search

○ Text feature ○ Image feature ● Compound

This interface can be divided into four parts.

- The text area at the top is for inputting query keywords in this panel.

- The "Visual Examples" panel is to create an example-based query based on the selection of user. The 'Add' button is used to add image examples into an image sample list, the elements of which are used to create a query based on image examples. The 'Delete' button is used to delete one or some of the image samples. The image samples are provided by TREC Video 2003.

- The "Select feature for Search" is used to select the type of features which will be used in the current search session.

- 'Search' button is used to run the query.

The Process of dealing with Data in this interface:

Step 1: Input query keywords in the top text area

Step 2: Select image samples for creating an image-based query

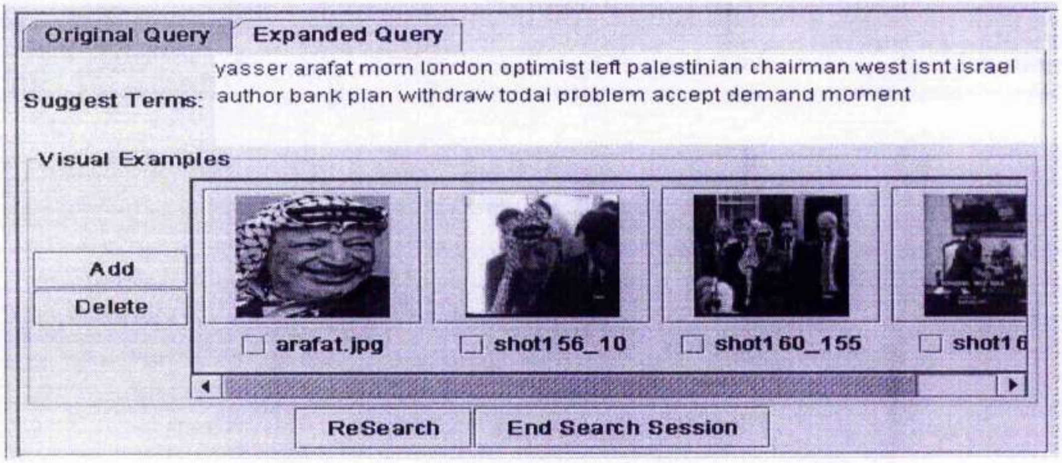
Step 3: Select an option of feature type which will be applied into the current search session. Choosing 'text feature' will use only textual features and choosing 'image feature' will use image features only. The 'Compound' selecting will use both features.

Step 4: Click 'Search' Button to perform retrieval.

4.3.1.2 Expanded Query Panel

The Expanded Query Panel is used to recommend a query based on terms, image samples, or the combination of terms and image samples (Figure 4.4). The details of expansion will be discussed later in the thesis.

Figure 4.4 Expanded Query Panel



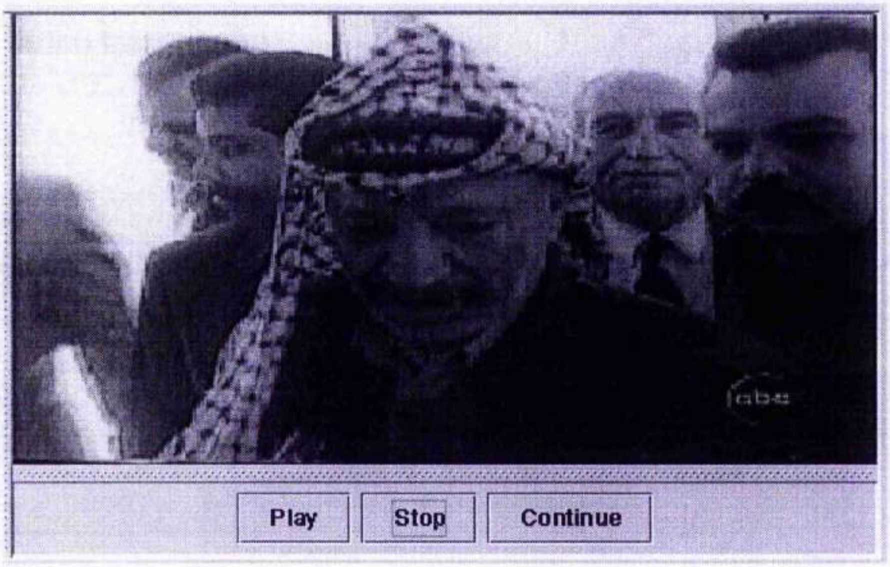
Suggest Terms: yasser arafat morn london optimist left palestinian chairman west isnt israel author bank plan withdraw todai problem accept demand moment

- This interface can be divided into three parts which have the similar functions as the Original Query Panel. However, it is only for the purpose of editing expanded query. The 'ReSearch' button is used to run the query which a searcher modified .

4.3.2 PlayBack Interface

The interface is divided into two parts, showed in Figure 4.5:

Figure 4.5 Playback Panel



1) The interface of play control:

a. "Play" Button: Play the current video shot

b. "Stop" Button: Stop and pause the play of videos

c. "Continue" Button: Continue playing video file without considering the limitation of the current shot boundary.

2) The interface for displaying the video stream of the current video shot.

4.3.3 Result Display Interface

The result of a query is displayed through a table model. The result table will show information about shots of the set of results, including key-frames and important words. Check box is used for explicitly marking the relevance of each video shot in the result set (Figure 4.6).

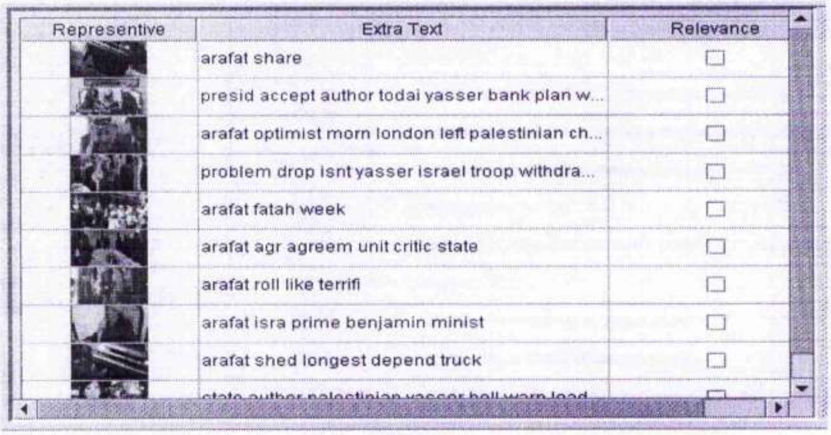
Function of result table:

1) Display the result of a query,

2) When users select one of shots in the result set, users can playback this shot from the start time to the end time displayed in the result table by using the interface for playing videos.

3) Mark relevance of shots:

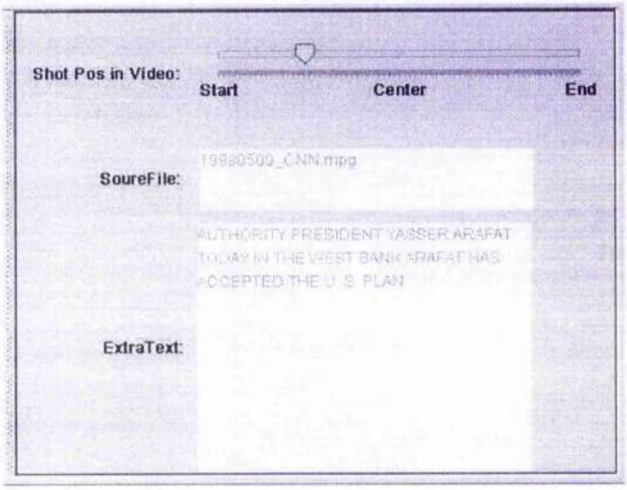
Figure 4.6 Display result panel for explicit feedback model



Representive	Extra Text	Relevance
	arafat share	□
	presid accept author todai yasser bank plan w...	□
	arafat optimist morn london left palestinian ch...	□
	problem drop isnt yasser israel troop withdra...	□
	arafat fatah week	□
	arafat agr agreem unit critic state	□
	arafat roll like terrifi	□
	arafat isra prime benjamin minist	□
	arafat shed longest depend truck	□
	state author palestinian yasser hall warn lead...	□

4.3.4 Showing Related Information Interface

Figure 4.7 Related Information Display Panel



Shot Pos in Video:

SoureFile: 19980500_CNN.mpg

ExtraText: AUTHORITY PRESIDENT YASSER ARAFAT TODAY IN THE WEST BANK ARAFAT HAS ACCEPTED THE U S PLAN

This interface is only an interface for showing the information of the current selected shot by users, which includes the following information:

- Shot Pos in Video: Display the position of the current selected shot in the video stream.
- SourceFile: Display the source file of the current selected shot
- Extra Text: Display the original ASR Result text of the currently selected shot.

4.4 Objectives of the System

The purposed of the system is to study the effect of interactive video retrieval schemes. We have developed a number of interactive retrieval models based on a set of implicit factors. These models are benchmarked using a simulation based strategy and is explained in the next chapter. The best performing model is used to build our system and is used for the experimental study.

Based on the results of the experiments in Chapter 3, we use 0.3 for weighting image features, and 0.7 for weighting textual features. Linear combination method is used for combing evidence.

Chapter 5

Simulation of Implicit Feedback Models

In this chapter, I will discuss the simulation of implicit feedback models. In the section 5.1, I proposed four implicit indicators based on a basic interactive video retrieval system--JIVRSystem introduced in Chapter 4. In the section 5.2, I will propose four implicit feedback models and the simulation of those implicit feedback models. The results and discussion will be presented in the final section.

5.1 User's actions and Implicit Factors

5.1.1 User's actions

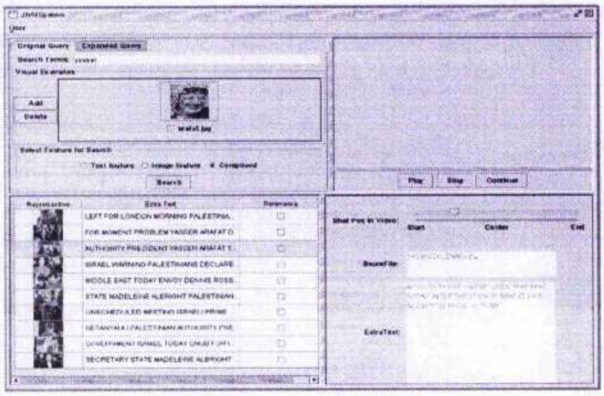
User's actions are different according to the interface support which a video retrieval system provides. However, basically, firstly user will construct an original query according to one's own information needs and run the original query, secondly user will browse the results returned by the video retrieval system. And then user will check the relevance of a retrieved video shot by clicking this video shot he or she is interested in, viewing the text-based summary or keyframe of this video shot, or playing this video shot.

5.1.2 Implicit factors

The idea is to refine the query based on the user basic interactions described above. We infer new query based on the cues inferred from user interactions. It is assumed that the following actions take place when a user is using a video retrieval system. Based on the user query, the video retrieval system presents a ranked list of shots along with textual and image snippets (thumbnail image and the text-based summary of this item -- in lower left panel described in Chapter 4). The user will browse the result set and will select a shot item which he or she is interested in. Here we assume that user actions correspond to their underlying need. This will result in displaying the details of the shots in lower right panel. If the user thinks that this item is relevant to his/her information need, he or she will play this item, and will see if this item is really relevant to his or her needs. It is assumed that if the item is relevant user will play it.

For studying the performance of implicit factors, I used a simulated search evaluation strategy and used TRECVID2003 topics 100-124 and took queries from the TRECVID2003 topic description. Query categorization and feature dependent weights were used from the previous chapter. I followed Yan's classification (Rong, Yang, and Hauptmann *et al.*, 2004) in this work. The appropriate features will be applied in the similarity measure. Different settings about the number of top ranked shots for the generation of relevance path in the simulation work are used in the work. For the queries where there are no relevant shots in the top N shots, the precision and recall are equal to zero.

Figure 5.1 Interface of the proposed system



Representative	Extra Text	Relevance
	LEFT FOR LONDON MORNING PALESTINIA...	□
	FOR MOMENT PROBLEM YASSER ARAFAT D...	□
	AUTHORITY PRESIDENT YASSER ARAFAT T...	□
	ISRAEL WARNING PALESTINIANS DECLARE...	□
	MIDDLE EAST TODAY ENVOY DENNIS ROSS...	□
	STATE MADELEINE ALBRIGHT PALESTINIAN...	□
	UNSCHEDULED MEETING ISRAELI PRIME...	□
	NETANYAHU PALESTINIAN AUTHORITY PRE...	□
	GOVERNMENT ISRAEL TODAY UNDER [illegible]	□
	SECRETARY STATE MADELEINE ALBRIGHT...	□

Based on this system, we propose the use of four different implicit factors in our video retrieval system: (1) Selecting a result item by moving the mouse over the item; (2) clicking the selected item; (3) view the key-frame, and text-based summary of a shot (4) playing a result item and/or Time of playing a result item.

5.2 Implicit Feedback Models

In this section, I propose the four implicit feedback models using the four implicit factors described in previous section

5.2.1 Binary Voting Model (BVM)

Binary Voting Model is a heuristic-based implicit feedback model (White *et al.* 2003). The objective is to identify features for refining the query from the documents viewed by the user. In its original implementation BVM is applied to a web search system and has been proven to be an effective method for improving the performance of it. It can be used to develop a retrieval model for video searching using the four implicit factors I proposed above. Though the general principle is the same as in White *et al.* 2003, we adapted it for video retrieval purposes.

In the case of video retrieval, a video shot is described by textual features which include many terms, and different kinds of image features extracted from the key-frame of it.

In the BVM video retrieval model, the four implicit factors are utilized to select new query terms and update an image query. The four implicit factors have the ability to indicate which video shot these four implicit factors derived from has the most indicativity.

When an implicit factor is used, the accessed video shot receives a "vote", and the terms appear in the shot will be given a weight. Image features of the key-frame of the accessed shots will be quantified by the indicativity values. The image query will be expanded with the weighted image features by computing the centroid of image features of the key-frames of the shots across the whole path if it is available. When it is not accessed, the corresponding accessed shot receives no vote. These votes accumulate across all viewed video shots.

It is asserted that the winning terms are those with the most votes. The assumption is that useful terms will be those contained in many video shots that a searcher accesses by various actions, the useful image features for expanding the image sample query will be the centroid of all corresponding key-frames of accessed video shots. The rationale behind this assertion that searchers will try to maximize the amount of relevant textual information and centralize the visual information they access during a search. The non-stopword terms that appear in the representation of those shots they view (and in similar contexts to their original query terms), and the centroid of all corresponding key-frames is the one that is potentially important to the searcher and may be useful for query modification.

5.2.1.1 Indicativity

In its original development, BVM model used indicativity weights (White *et al.* 2003). We also used the same approach, adapted and followed it for our retrieval scenario. The weights assigned to the four implicit factors actually present the indicativity of the video shots the four implicit factors work on. Different implicit factors vary in the ability to indicate a searcher's information need. It is an assumption that the contribution the implicit factor--playing a video shot makes to the system's understanding of which shot is relevant to be more than other three implicit factors. The action of viewing the text summary and key-frame of a video shot is more indicative than the action of clicking a video shot. The indicativity of movement of mouse over a video shot is the least among these four implicit factors.

In the video retrieval system I developed, Indicative weights for various implicit factors are chosen according to the above empirical assumption. The weights chosen are:

1. When user move mouse over a shot item in the result panel, system will highlight the shot item, a weight of 0.1 is given

2. When user clicks a shot item in the result panel, a weight of 0.2 is given for the action

of click.

3. After user click a shot item in the result panel, one views the key-frame and text-based summary of the currently selected shot item, a heuristic weight of 0.2 is given for the action of viewing the key-frame and text-based summary.

4. When user plays a shot item in the result panel, a heuristic weight of 0.5 is given for the fourth implicit factor.

Here all four implicit factors proposed are ordinal. The action of moving mouse over a video shot will happen first, and if the searcher is interested in the current video shot below mouse, the second action is to click the video shot. The third is to view the text summary and the image of the corresponding key-frame of the clicked video shot. The final action is to play the shot for completely browsing the video shot.

For example, when a searcher clicks a video shot in the result panel, he or she views the key-frame and text-based summary of the currently selected shot item, a heuristic weight of 0.2 is given for the action of viewing the key-frame and text-based summary.

This means all terms appear in the clicked video shot will receive a weight of 0.2, because logically, if a searcher clicks a video shot, the first action of the searcher is to move mouse over the video shot, and then the action of clicking happens. Therefore, the weight of 0.1 assigned for the implicit factor-mouse over a video shot should be added with the weight of 0.2 assigned for the action of clicking a video shot. Therefore, all terms appear in the clicked video shot will receive a weight of 0.3 (0.1+0.2), which is also used to weight the corresponding image features which describe the clicked video shot. First, we will explain the relevance path used, and then how the terms are weighted will be introduced.

5.2.1.2 Relevant Path

In this study, simulation paths are extracted only from relevant video shots which are retrieved from the top N = 10, 30, 50, 100 results for each of the 25 TRECVID2003 topics used as queries. These results can contain both relevant and non-relevant video shots. However, for some search topics, there are no relevant documents in the top N (10, 30, 50, 100) results, making the execution of the scenario problematic. In this case, both precision and recall are equal to 0.0, but it will not result in important impact on the search effectiveness, because I average the search effectiveness of the 25 search topics.

The following show the possible paths (Relevance Path). A searcher can traverse within the representations of a video shot. After the user play a video shot, we create a new query:

Table 5.1 Possible Relevance Path

		User's Behaviors			
Move mouse over a video shot	Click a video shot	View text summary and key-frame	Play a video shot		Total Paths
1	1	1	1		1
1	1	1			1
1	1		1		1
1	1				1

1					1

For example, when a searcher does all four actions (first row of Table 5.1) there are 1 × 1 ×1× 1 = 1 possible paths. The final column shows the total for each possible route. There are 5 possible relevance paths for each video shot. If all top N (N= 10, 30, 50, 100) video shots are used, there are N×5 (5× 10, 5×30,5×50,5×100) possible relevance paths per search topic, but in real life, it is possible that , after a searcher view the textual summary and key-frame of a video shot, the searcher may access other video shots which have been accessed according to the above routes. Therefore, it should be assumed that a searcher only accesses each video shot in the N top ranked video shots one time through one of the routes above. This strategy is same as what used in White *et al.*, 2005.

5.2.1.3 Term weighting

For weighting term we follow the model as described in (White *et al.* 2003). The BVM model is a simple approach to a potentially complex problem. The terms with most votes are those that are taken to best describe the information viewed by the searcher (i.e., those terms that are present most often across all viewed shots) and can therefore be used to approximate searcher interests. Of course, searchers may view irrelevant information as they search. In general however, their interaction decisions are guided by a desire to maximise the amount of relevant information they view.

The textual feature of each video shot is represented by a vector of length n; where n is the total number of unique non-stopword, stemmed terms extracted from ASR results. All terms are candidates in the voting process.

To weight terms, a shot × term matrix, shown in Figure 5.2, (s+1)× n is constructed, where s is the number of documents for which the searcher has visited. Each row in the matrix represents all n terms in the vocabulary [i.e., $(t_{k1}, t_{k2}, \ldots, t_{kn})$ where k is the row number], and each term has a weight. An additional row is included for the query.

Figure 5.2. Shots × Term matrix.

$$\begin{array}{ccccccc} & & t_1 & t_2 & \cdots & t_n & \\ Q_0 & [& t_{01} & t_{02} & \cdots & t_{0n} &] \\ S_1 & [& t_{11} & t_{12} & \cdots & t_{1n} &] \\ S_2 & [& t_{21} & t_{22} & \cdots & t_{2n} &] \\ & [& \cdots & \cdots & \cdots & \cdots &] \\ S_s & [& t_{s1} & t_{s2} & \cdots & t_{sn} &] \end{array}$$

Query terms are initially assigned a weight of one if they are included in the query and zero if not. Example 5.1 illustrates the operation of the Binary Voting Model.

Example 5.1: Simple Updating

If one assumes that there are only 10 terms in the vocabulary space in the collection and that the original query (*Q*0) contains *t*3 and *t*7, the document × term matrix initially looks like:

$$\begin{array}{ccccccccccccc} & & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 & t_{10} & \\ Q_0 & [& 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 &] \end{array}$$

Each row in the matrix is normalised to give each term a value in the range [0, 1] and make the values sum to one which ensures that the query terms are not weighted too highly in the shots × term matrix. This is important when the model is *replacing* query terms; a high query term weight would lessen the chances of other terms being chosen. The matrix now looks like:

$$\begin{array}{cccccccccccc} & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 & t_{10} \\ Q_0 & [\;0 & 0 & .5 & 0 & 0 & 0 & .5 & 0 & 0 & 0\;] \end{array}$$

Each document representation is regarded as a source of terms, and the act of viewing a representation as an implicit indication of relevance. When a searcher visits the first representation for a video shot a new row is added to the document × term matrix. This row is a vector of length n, where n is the size of the vocabulary and all entries are initially set to 0. If a term occurs in a representation, no matter how many times, it is assigned a weight, w_t, which is based on the representation that contains the term.

This weight for each term is *added* to the appropriate term/document entry in the matrix. Weighting terms is therefore a *cumulative* process; the weights calculated for a term in one representation are added to the weights calculated for the preceding steps in the relevance path. The Binary Voting Model calculates weights on a per video shot basis (i.e., within video shot). There are different sets of weights for each video shot and these weights correspond to a row in the shot × term matrix. The total score for a term in a shot is computed by:

$$W_{t,S} = \sum_{a=1}^{p} W_a * W_t \qquad (5.1)$$

Where p is the number of steps taken by the user, a is the action of the searcher W_a is the heuristic weight for the action a (as explained above) and W_t is the binary weight of term in a representation.

Example 5.1: Simple Updating (continued)

When a searcher follows a relevance path of implicit factors, the model updates the weights in the shot × matrix after each step. How the term weights are updated as a path from the action of mouse movement, to the action of viewing text summary and the image of the key-frame is traversed, is following.

It is assumed that S_6 [t_3 t_4 t_5 t_6 t_7] where S_6 is the ID of the shot, $t_3 \dots t_7$ are the terms appear in the S_6.

When a searcher move mouse over Shot S_6, the updated list of terms of S_6 goes as follows:

$$\begin{array}{cccccccccccccc} & & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 & t_{10} & \\ Q_0 & [& 0 & 0 & .5 & 0 & 0 & 0 & .5 & 0 & 0 & 0 &] \\ S_6 & [& 0 & 0 & .1 & .1 & .1 & .1 & .1 & 0 & 0 & 0 &] \end{array}$$

When a searcher clicks the Shot S_6, the updated a list of terms of S_6 goes as follows:

$$\begin{array}{cccccccccccccc} & & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 & t_{10} & \\ Q_0 & [& 0 & 0 & .5 & 0 & 0 & 0 & .5 & 0 & 0 & 0 &] \\ S_6 & [& 0 & 0 & .3 & .3 & .3 & .3 & .3 & 0 & 0 & 0 &] \end{array}$$

When a searcher views the text summary and key-frame of the clicked Shot S_6, the updated a list of terms of S_6 goes as follows:

$$\begin{array}{cccccccccccccc} & & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 & t_{10}' & \\ Q_0 & [& 0 & 0 & .5 & 0 & 0 & 0 & .5 & 0 & 0 & 0 &] \\ S_6 & [& 0 & 0 & .5 & .5 & .5 & .5 & .5 & 0 & 0 & 0 &] \end{array}$$

After the above three actions, the list of candidate terms goes as follows:

$$\begin{array}{cccccccccccccc} & & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 & t_{10} & \\ C_6 & [& 0 & 0 & 1.0 & .6 & .6 & .6 & 1.0 & 0 & 0 & 0 &] \end{array}$$

Where C_6 is the list of weights of all candidate terms based on shot S_6.

When the searcher moves mouse over the other shot S_{10} $[\ t_1 \quad t_2 \quad t_4 \quad t_6 \quad t_9\ ]$

The updated Shot x terms matrix is following:

$$\begin{array}{lcccccccccccc}
 & & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 & t_{10} & \\
Q_0 & [& 0 & 0 & .5 & 0 & 0 & 0 & .5 & 0 & 0 & 0 &] \\
S_6 & [& 0 & 0 & .5 & .5 & .5 & .5 & .5 & 0 & 0 & 0 &] \\
S_{10} & [& .1 & .1 & 0 & .1 & 0 & .1 & .1 & 0 & 0 & 0 &] \\
C_{6,10} & [& .1 & .1 & 1.0 & .6 & .5 & .6 & 1.1 & 0 & 0 & 0 &]
\end{array}$$

Where $C_{6,10}$ is the list of weights of all candidate terms based on shots S_{10} and S_6

At this time, a normalized function will be applied to make the sum of weights of all candidate terms equal to 1.0. The final vector of weights of all candidate terms goes as follows;

$$\begin{array}{lcccccccccccc}
 & & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 & t_{10} & \\
C_6 & [& .025 & .025 & .25 & .15 & .125 & .15 & .275 & 0 & 0 & 0 &]
\end{array}$$

I made an assumption that, once searcher finishes the action of playing a shot, the system will generate a new query and issue this new query. The terms with overall votes are the candidate terms for query expansion. Here, 20 top ranking terms are selected for the new query.

5.2.1.4 Query Modification for textual feature

In the matrix, only the query terms corresponding to shots accessed by the searcher will have a score greater than zero. The set of terms in the accessed video shots is potentially helpful for query modification.

After accessing the implicit factor--the action of playing a shot, which allows the model to gather sufficient implicit evidence from searcher interaction, a new query is constructed. It is possible for a relevance path to access different implicit factors of different shots.

To compute the new query, the framework calculates the total score for each term across all shots (i.e., down each column in the shots × term matrix) and then the normalized scores will be computed. This gives a normalize score for each term in the vocabulary. The terms are then ranked by it. A high normalized score implies the term has high indicative weights across the shots viewed. The top 20 ranked terms are used modify the query. According to White *et al.*, 2005, there are two main ways: *query expansion* and *query replacement*.

Query expansion – The top N terms chosen by the Binary Voting Model are appended to the original terms chosen by the searcher.

Query replacement – It is possible that the new query may not contain the searcher's original query terms; this would be a form of query replacement as the estimated information need has changed sufficiently to warrant the original query being completely replaced.

In this study, I adopted the second way--Query replacement for my simulation work based on an assumption that the new terms are relevant to user's information needs and it has low possibility that original terms is removed from new query, though it is possible that new query may not contain the searcher's original query terms, other new terms have enough relevant information for retrieving more documents relevant to user's information needs.

5.2.1.5 Image Feature Weighting

The formation of image query is not trivial. Images are displayed as key-frames in the result list, and also enlarged versions appear in another panel. A new image query is the weighted centroid of images along the path.

A path is defined as a number of representations of a particular document a user is viewing. The following formula is used to weight one vector of image feature for a query:

$$Q'_{image} = Q_{image} + \frac{1}{n}\sum_{a=1}^{p} F_{image,a} * W_a \qquad (5.2)$$

where p is the number of steps taken, Q'_{image} is the new image feature vector, $F_{image,a}$ is the image feature corresponding to an action, W_a is the heuristic weight for the action, a is the action of searcher, Q_{image} is the original query vector for a given feature, and n is the number of the images in a path. This is a replicated for all the features.

The system will update the weights of candidate terms and image features once user start to operate the action of playing a shot. Subsequently, the system will issue a new query and the new set of results will be presented to the user. The system will rank every candidate terms, and select the top 20 terms as the query terms which will be updated in each stage. The reason to choose top 20 terms is based on experimental evidence in which we tried the use of top 10, 20, 50, 100 terms. Using the top 20 terms, we got the balance between the effectiveness, the speed and the cost of system source.

5.2.2 Binary Voting Model Variant

In the above model, the action of playing a video shot is the most indicative implicit factor. But, in the variant of BVM model I propose in this subsection, the time of playing a video shot is considered instead of the implicit factor—playing a shot.

In this case, the hypothesis we assume is that the time of playing a shot is a very good interest indicator and can be used to infer searchers' information needs. Therefore, the time of playing a video shot is used to substitute the action of playing. The approach incorporates this implicit factor—time of playing a video shot.

In this particular case instead of using a weight of 0.5, I used the following the function to arrive at a weight:

$$W_i = T_{time\_playing} / T_{time\_shot} \qquad (5.3)$$

Where W_i is the weight of a feature in a shot which is selected and played by searchers; $T_{time\_playing}$ is the time of playing a video shot, T_{time_shot} is the duration of the current shot being played. We hope the arrived weight will be above 1.0 which is used to represent higher relevance. The following function was used to weight the duration of playing in the real systems and then we normalize this when a searcher runs the new query:

$$W_i = 1.0 + random() \qquad (5.4)$$

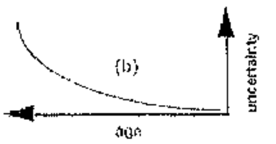
Where random() is a Java function which is used to generate a float number (0.0≤random()<1.0).

5.2.3 Ostensive Binary Voting Model

Based on the Binary voting model, I developed the ostensive binary voting model, which is to assign these ordinal implicit factors with ostensive relevance profile (Campbell and Van Rijsbergen *et al.*, 1996) instead of the pre-defined heuristic weights for implicit factors.

The profile is of a decelerating increase in uncertainty with age. It means that the most recent evidence has the lowest attached uncertainty and therefore will have the most influence on the weighting. Here, all evidence plays a part in the ostensive definition; nevertheless, the most recent judgment will play the greatest role (Campbell and Van Rijsbergen et al., 1996).

Figure 5.3 Decelerating increase in uncertainty with age



The following function is defined for generating the related ostensive profile:

$$W_i = -2^{-i} \qquad (5.5)$$

Where W_i is the weight of i_{th} relevance node, which is used to weight order i_{th} implicit factor; i is the index of ordered implicit factors. A process will be applied to move negative values to be positive, which uses the following functions:

$$W_i' = W_i + abs(\min(W_i)) \quad (5.6)$$

Where W_i' is the new weight of term i, *abs*() is a function which is used to compute the absolute value, min() is a function which is used to compute the minimum value of all W_i. Then a normalized process will be applied to make sum of weights of all candidate terms equal to 1.0.

5.2.4 Pure Ostensive Model

In this model, I do not consider the four implicit factors I propose, but only use the basic approach of ostensive model (Campbell and Van Rijsbergen *et al.*, 1996), which only considers the behavior of double clicks on a viewed shot of a result set as an effective implicit factor, and use ostensive profile to weight it. Therefore, based on this model, video retrieval system only catches the user's behaviour of double click on an item of the result of a previous search operation. The candidate terms are derived from the item double-clicked by the user. The weights of these candidate terms are from the ostensive profile (I adopt ostensive profile of decelerating increase in uncertainty with age.) proposed. A normalizing process is also applied at the end.

Example **5.1 (continued)**:

It is assumed that S_6 [t_3 t_4 t_5 t_6 t_7] where S_6 is the ID of the video shot, $t_3 \ldots t_7$ are the terms appear in the S_6.

When a user double clicks a video shot item S_6, in the result set, a weight of 1.0 will be given by Ostensive profile of decelerating increase in uncertainty with age. The updated list of terms of S_6 goes as follows:

$$\begin{array}{cccccccccccccc}
 & & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 & t_{10} & \\
Q_0 & [& 0 & 0 & .5 & 0 & 0 & 0 & .5 & 0 & 0 & 0 &] \\
S_6 & [& 0 & 0 & w_3 \times 1.0 & w_4 \times 1.0 & w_5 \times 1.0 & w_6 \times 1.0 & w_7 \times 1.0 & 0 & 0 & 0 &]
\end{array}$$

When a searcher double-clicks the Shot S_{10} [t_1 t_2 t_4 t_6 t_9], the updated a list of terms of S_{10} goes as follows:

$$\begin{array}{cccccccccccccc}
 & & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 & t_{10} & \\
Q_0 & [& 0 & 0 & .5 & 0 & 0 & 0 & .5 & 0 & 0 & 0 &] \\
S_6 & [& 0 & 0 & w_3 \times 1.0 & w_4 \times 1.0 & w_5 \times 1.0 & w_6 \times 1.0 & w_7 \times 1.0 & 0 & 0 & 0 &] \\
S_{10} & [& w'_1 \times 0.5 & w'_2 \times 0.5 & 0 & w'_4 \times 0.5 & 0 & w'_6 \times 0.5 & 0 & 0 & w'_9 \times 0.5 & 0 &] \\
C_{6,10} & [& w'_1 \times 0.5 & w'_2 \times 0.5 & w_3 \times 1.0 & w_4 \times 1.0 + w'_4 \times 0.5 & w_5 \times 1.0 & w_6 \times 1.0 + w'_6 \times 0.5 & w_7 \times 1.0 & 0 & w'_9 \times 0.5 & 0 &]
\end{array}$$

Where $C_{6,10}$ is the list of weights of all candidate terms after double-clicking shots S_{10} and S_6. After each process of weighting, a process of normalization describe in the previous section, will be applied to make the sum of weights of all candidate terms equal to 1.0.

5.2.5 Simulation based evaluation methodology

For benchmarking the effectiveness of various models, we follow the approach proposed in White *et al.*, 2005. Real searcher would typically follow a series of related relevance path in a rational way, viewing only the most useful or interesting. In this study, the actions I try to simulate are the four implicit factors. The simulation approach is similar to the methodology described in White *et al.* 2003 which has been proven to be regarded as a feasible methodology for evaluating the effectiveness of RF technology, because Simulation-centric methods are less time consuming and costly than experiments with human subjects, and allow the comparison of IR techniques in different retrieval scenarios, and maintain control over environmental and situational variables.

In this section, I will introduce simulation based evaluation methodology being used in this study. Firstly, I present the context of the simulation, which include system corpus and search topics, Secondly, the Evaluation Procedure will be described, and the experimental results will be proposed and analyzed. Finally, a conclusion will be made.

5.2.5.1 System, Corpus and Topics for Simulation

The video collection defined by TRECVID is really particular for the research on video retrieval. The test collection of this complete video collection defined by TRECVID2003 totally includes 121 video files of ABC World News Tonight and CNN Headline News, which includes 35220 video shots.

TREC topics 100-124 defined by TRECVID2003 were used and the query was taken from the short descriptions of the search tasks. For each query, I will use the top 10, 30, 50, 100 video shots for generating relevance paths for use in the simulation respectively.

The number and nature of relevance paths chosen for the simulation is dependent on the simulation strategy employed, i.e., the interact model of simulating searchers, the selection of relevance paths. There are three main strategies for selecting relevance paths.

1. All video shots in the top N = 10, 30, 50, 100 ranked video shots are used to create relevance paths.

2. All relevant shots in the top N = 10, 30, 50, 100 ranked video shots to a search topic are used to generate relevance paths.

3. All non-relevant shots in the top N = 10, 30, 50, 100 ranked video shots are used to generate relevance paths.

5.2.5.2 Evaluation Procedure

The simulation creates a set of relevance paths for all relevant video shots in the top-ranked documents retrieved for each topic. The number of depends on the simulation strategy employed.

After each iteration the effect on search effectiveness was monitored. The precision is a measure of search effectiveness. In this study, I define the end of an iteration as the end of the finishing playing the current selected shot. I compute the precision and recall at iterations 3, 5, 10, 15 and 20 and record them. Repeating the above process for eight times is for the purpose of obtaining the average precision and recall, which is to avoid the possibility of a very good or very bad performance at a given trial as well as Ryen's method (White *et al.*, 2005).

In this study, if searchers finish the action of playing a shot, the system will generate a new query and issue this new query. That means that the effect of each implicit factor will continue to being accumulated until searchers finish the action of playing. We repeat each feedback iteration for 8 times and obtain the average performance of all TRECVID2003 search topics.

Using similar procedure as White *et al.*, 2005, the following procedure is used for each topic with each model:

I. Use JIVRSystem (I have introduce the basic structure of that system) to retrieve document set in response to each search topic.

II. Identify relevant or non-relevant documents in the top N (N =10, 30, 50, 100) retrieved video shots, depending on the experimental run and store in set s.

III. Create and store all potential relevance paths for each relevant video shot in s.

IV. Choose relevance paths as suggested by the simulation strategy, setting m to the number chosen. A random number generator is used where appropriate in selecting random paths.

V. For each of the m relevance paths/video shots:

a. Weight terms and image features in path/video shots with chosen model and rank terms based on heuristic weights pre-defined by myself.

b. Use top-ranked 20 terms and weighted image features to expand original query.

c. Use new query which may include textual query and a query by an image example to retrieve new set of documents.

d. Compute new precision values.

VI. Repeat from II to V for eight times,

VII. Compute the average precision and recall of results of eight times

To better represent a searcher exploring the information space, all subsequent retrievals were not only to test the search effectiveness of the new queries and were used to generate relevance paths for next feedback iteration.

5.2.6 Results

The study was conducted to benchmark a variety of implicit feedback models using the four implicit factors in the context of interactive video retrieval. In this section, I present results of the study. I focus on results concerning search effectiveness. I use BVM, BVM_OS, BVM_TIME, and POS to refer to the Binary Voting Model, the variant of Binary Voting Model which adopts ostensive profile to weight the four implicit factors, the variant of Binary Voting Model which weights the implicit factor-time of playing a video shot instead of the implicit factor-playing a video shot, and Pure Ostensive Model, respectively. All of these abbreviative words (BVM, BVM_OS, BVM_TIME, and POS) will be used in the following sections.

Search effectiveness

In this study, I use the relevant subset strategy which uses a set of relevance paths taken from the relevant shot video shots from top-ranked video shots. This strategy assumes that all the video shots a searcher views is relevant to search topics.

Figure 5.4 Precision accorss 8 runs at the top-ranked 30 video shots

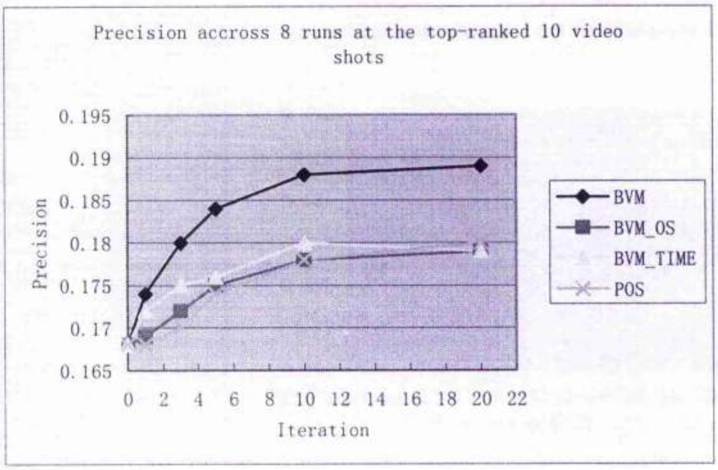


Figure 5.5 Precision accorss 8 runs at the top-ranked 30 video shots

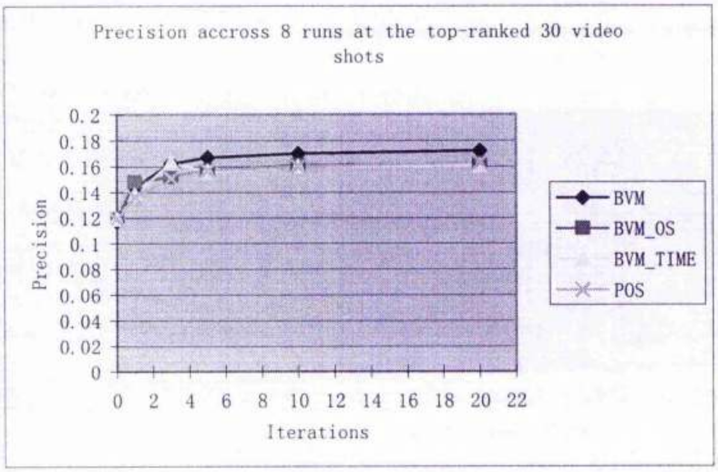
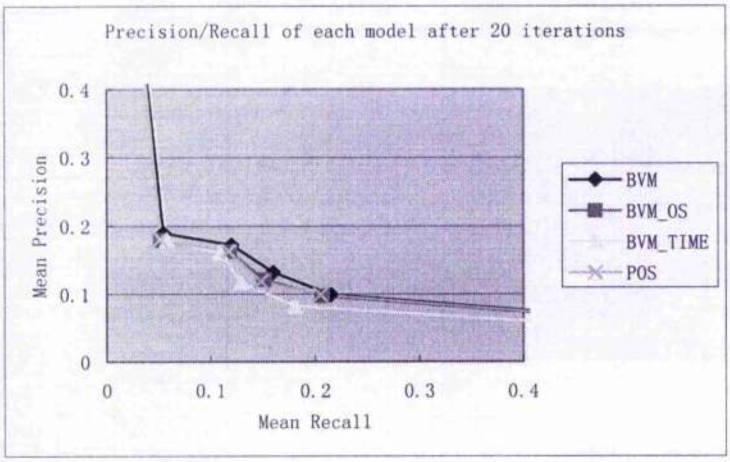


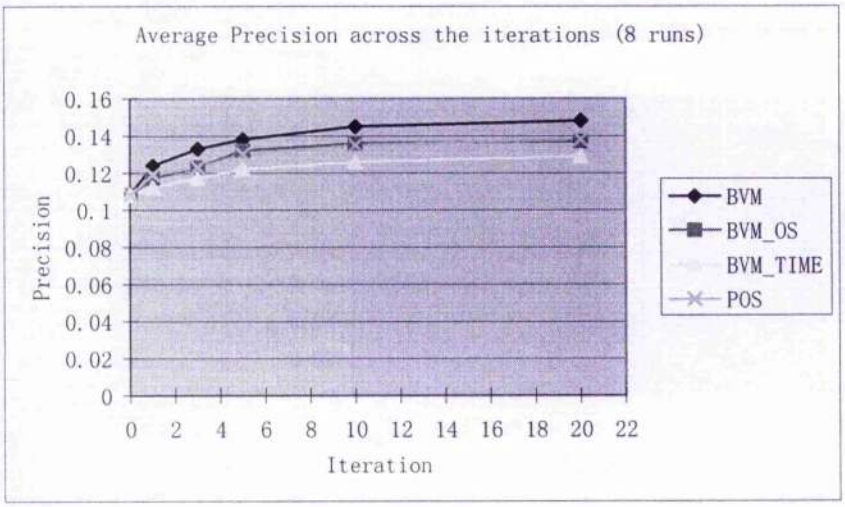
Figure 5.4, 5.5 show that BVM has the best performance at the top-ranked 10, 30 video shots. Pure ostensive model and BVM_OS model has almost same performance and has better performance than the BVM_TIME model, the BVM_TIME model perform worst than other three models.

Figure 5.6 Precision/Recall of each model after 20 iterations.



According to Figure 5.6, after 20 iterations, the BVM_TIME model still perform poorly, the pure Binary voting model remains the best performance. The BVM_OS model and the Pure Ostensive Model have similar performance and better than the BVM_TIME model. BVM model is a little bit better than other three models.

Figure 5.7 Average Precision across the iterations (8 runs)



According to Figure 5.7, it is an obvious conclusion that the overall performance of the Pure Binary Voting Model is better than other three models with large increases inside the first five iterations, and Pure Ostensive Model and BVM_OS model have similar performance of each other, and the BVM_TIME model performance more poorly than other three models. The Binary Voting Model is quick respond to implicit relevance information, with more marginal increases. There is a steady increase until around 10 iterations where precision levels out though the marginal effects of all models appears slight.

Table 5.2—5.5 illustrates the marginal difference more clearly than Figure 5.4-Figure 5.6

Table 5.2 Percentage change in Precision at the point of 10 documents across the number of iterations

	0		1		3		5		10		20	
BVM	16.8	---	**17.4**	**0.6**	**18**	**0.6**	**18.4**	**0.4**	**18.8**	**0.4**	**18.9**	0.1
BVM_OS	16.8	---	16.9	0.1	17.2	0.3	17.5	0.3	17.8	0.3	17.9	0
BVM_TIME	16.8	---	16.9	0.1	17.1	0.2	17.3	0.2	17.7	0.4	17.9	**0.2**
POS	16.8	---	16.8	0	17	0.2	17.4	**0.4**	17.8	**0.4**	18	**0.2**

Table 5.3 Percentage change in Precision at the point of 30 documents across the number of iterations

	0		1		3		5		10		20	
BVM	11.9	---	**14.4**	**2.5**	**16.2**	**1.8**	**16.7**	0.5	**17**	0.3	**17.2**	0.2
BVM_OS	11.9	---	12.5	1.6	13.2	0.7	14.3	1.1	15.1	**0.8**	15.2	0.1
BVM_TIME	11.9	---	12.1	0.2	12.9	0.8	13.4	0.5	14	0.6	14.5	**0.5**
POS	11.9	---	13.6	1.7	14	0.4	14.6	0.6	15.2	0.6	15.3	0.1

Table 5.4 Percentage change in Precision at the point of 50 documents across the number of iterations

	0		1		3		5		10		20	
BVM	8.56	---	**10.4**	**1.84**	**11.6**	**1.2**	**12.1**	0.5	**12.8**	**0.7**	**13.2**	**0.4**
BVM_OS	8.56	---	10.1	1.54	11.2	1.1	11.6	0.4	11.9	0.3	12.1	0.2
BVM_TIME	8.56	---	10	1.44	10.4	0.4	10.9	0.5	11.3	4	11.7	**0.4**
POS	8.56	---	10.2	1.64	10.8	0.6	11.7	**0.9**	11.6	-0.1	11.9	0.3

Table 5.5 Percentage change in Precision at the point of 100 documents across

	0		1		3		5		10		20	
BVM	5.92	---	**7.22**	**1.3**	7.55	0.33	8.01	0.46	**9.51**	**1.5**	**9.92**	0.41
BVM_OS	5.92	---	7.18	1.26	**7.67**	0.49	**9.33**	**1.66**	9.43	0.1	9.71	0.28
BVM_TIME	5.92	---	6.08	0.16	6.52	0.44	7.11	0.59	7.53	0.42	8.24	**0.71**
POS	5.92	---	6.03	0.11	7.57	**1.54**	8.26	0.69	9.32	1.06	9.68	0.36

Table 5.6 Percentage change in precision per iteration. Overall Change in first column, marginal change in second shaded column. Highest percentage in each column iteration

	0		1		3		5		10		20	
BVM	10.8	---	**12.4**	**1.6**	**13.3**	**0.9**	**13.8**	0.5	**14.5**	**0.7**	**14.8**	**0.3**
BVM_OS	10.8	---	11.7	0.9	12.3	0.6	13.2	**0.9**	13.6	0.4	13.7	**0.1**
BVM_TIME	10.8	---	11.2	0.4	11.7	0.5	12.2	0.5	12.6	0.4	12.9	**0.3**
POS	10.8	---	11.9	1.1	12.3	0.4	13	0.7	13.5	**0.5**	13.8	**0.3**

Table '5.2,5.3,5.4,5.5' which show the precision at the point of the top 10, 30, 50, 100 top-ranked video shots indicates that the largest increase from pure Binary Voting model, though the marginal effects of all models appear slight. The variant of Binary Voting

Model which gives a weight for the time of playing based on the length of playing perform poorly, although still lead to small overall increase in precision over baseline. Performance of both of the variant of Binary Voting Model which uses the weighting strategy--ostensive profile and the pure ostensive model is over baseline, and is quite similar.

In addition, Friedman Rank Sum Test was used to test the significant difference among the performance of these four models. The results (N = 5, Chi-Square = 11.659, df = 3, p = 0.009 < 0.05) for the top ranked 10 video shots, (N = 5, Chi-Sqaure = 15.00, df =3, p = 0.002 < 0.05) for the top ranked 30 video shots, (N = 5, Chi-square = 13.56, df = 3, Asymp.sig. = 0.004 < 0.05) for the top ranked 50 and video shots (N = 5, Chi-square = 9.24, df = 3, Asymp.sig. = 0.026 < 0.05) for the top ranked 100 video shots show that there is significant different among these models at the top-ranked 10, 30, 50, 100 video shots.

5.2.7 Discussion

The implicit feedback models evaluated in the study all increased search effectiveness through query modification. However, the pure Binary Voting Model performs particularly well; BVM_OS Model and Pure Ostensive Model have the similar performance; the BVM_TIME model performs really poorly. From the aspect of marginal effects, pure Binary Voting Model also has the largest marginal effect, the BVM_OS Model and Pure Ostensive Model have the almost similar marginal effect, the marginal effect of BVM_TIME Model is a little bit lower than Pure Ostensive Model and BVM_OS Model, but much lower than the Pure Binary Voting Model.

The Binary Voting Model selects terms based only on the implicit factors accessed by the searcher in the context of interactive video searching system and appropriately weighting corresponding image features of digital video. The lists of potential terms offered stagnates after 10 iterations, the effect of the scoring is cumulative, the high-scoring, high-occurrence terms, obtain a higher score after only a few initial paths and cannot be succeeded by lower-ranked terms in later paths. This often means that the same query is presented in iterations 10 and 20. In the study of White *et al.*, 2005, this effect has noticed as well. The findings of the study show that the Binary Voting Models is able to perform more effectively than the baselines when all the paths presented to them are from relevant video shots. For almost all iterations on all models, the models appear to reach a point of saturation at around 10 paths, where the benefits of showing 10 more paths (i.e., going to iteration 20) are only very slight and are perhaps outweighed by the costs of further interaction, because the marginal effect increases in precision as more relevant information is presented. A possible reason is that a new injection of different information may become needed because the relevance information reaches a saturation point. For example, explicit involvement may be an effective relevance information source.

Chapter 6

Evaluation

In Chapter 5 a heuristic approach towards implicit feedback retrieval was described. This approach uses searcher interaction with video shots to generate new query statements. The part concluded with a simulation-based evaluation of different candidate implicit feedback models.

The Binary Voting Model performed best and was therefore selected for developing an interface of video retrieval system. The experiment shows the effectiveness of different implicit feedback methods based on the particular factors proposed. Unlike the tests carried out in Chapter 5, this experiment involves human participants, and evaluates usefulness and effectiveness of the interface for a video retrieval system developed.

6.1 Introduction

In the previous chapter, the simulation-based study tested how well implicit feedback models improved search effectiveness. The study found that the Binary Voting Model outperformed the other models tested. In this chapter, the experiment also evaluates the form of interface support for presenting textual query and image sample based queries. In the interface studied, the amount of control searchers have over creating and expanding queries, and making search decisions is varied. The chapter begins by describing the user study, and then further describes the experimental methodology. Finally, results of this user study will be described.

6.2 User study

The aim of this user experiment is to evaluate various interface components such as a tool for suggesting terms and images and how much control users need. A prototype system developed based on the Binary Voting Model is used for this purpose. The goal of the study is to evaluate the interface support mechanisms and the effectiveness of the heuristic-based implicit feedback framework from users' perspective.

6.2.1 Experimental Hypotheses

Our experimental hypotheses are the following:

i) A combination system of implicit and explicit features is better than the system based on explicit feature only for video retrieval

ii) During a search session, the user's actions (e.g. playing a video shot, browsing video, or seeing related information of one video shot) in a video retrieval system can be captured and used as an indicator which shows the relevance of shots.

iii) The form of recommending a search query based on terms and image samples are comfortable and useful for participants.

6.2.2 Participants

24 experimental participants were recruited. The experimental participants were mainly staff and undergraduate and postgraduate students at University of Glasgow. Participants were paid £10 for participating. The study uses a within-subjects experimental design meaning that subjects used all experimental systems. A Greco-Latin square based design is used to control subjects' learning effects between systems (Tague et al., 1992).

6.2.3 System

Two versions of the JIVRSystem have been developed in order to compare the performance of different interactive video systems— one (System 1 or S1) is based on explicit feedback features, the other (System 2 or S2) is based on the combination of explicit and implicit features These two systems use the same index extracted from Automatic Speech Recognition, Close-caption and key-frames. Low-level image features are extracted based on global color histogram and Cooccurrence algorithms (Sonka, Hlavac and Boyle *et al.*, 1998). The two systems share a keyword-based and image-based interface, and the method of obtaining initial result set. However different mechanisms are used for reformulating queries for iterative search.

6.2.4 Document Collection

For the purpose of the experiment I employed the video collection recommended by TREC Video 2003 and described in Chapter 3. The video collection includes ABC World News Tonight and CNN Headline News recorded by the Linguistic Data Consortium from 21st April 1998 to 24th July 2001 the number of shots of which is over 60,000. The information about the boundaries of shots is described by a series of XML-based files, provided by LIMSI (LIMSI), which contains the ASR results text.

6.2.5 Search Tasks

The 25 search topics are defined by TRECVID2003. Based on categorization of search topics of Yan (Rong, Yang, and Hauptmann, 2004), I selected one search topic from each category. Four search tasks are used to test the usefulness and effectiveness of these two systems. Each of those three different search tasks belongs to a different category proposed by Yan (Rong, Yang, and Hauptmann, 2004).

The search topics selected are following:

Person:

0103-- Find shots of Yasser Arafat

Specific Object:

0106-- Find shots of the Tomb of the Unknown Soldier at Arlington National Cemetery

General Object:

0109-- Find shots of one or more tanks

Scene:

0117-- Find shots of one or more groups of people, a crowd, walking in an urban environment (for example with streets, traffic, and/or buildings).

These tasks are provided to the user using a simulated search situation (Borlund and Ingwersen *et al.*, 1997). The simulated task situation and background information is provided (Please see Appendix D).

6.2.6 Search Task Allocation

I use Greco-Latin Square Design (Recommend by TRECVID2003) for user experiments. In the Square, I use U-1, U-2,......U-24 refer to 24 subjects and use T1, T2, T3 and T4 refer to the four search tasks. The search task distribution is shown in Table 6.1.

Table 6.1 Search Task Distribution

U-1	S1, T1	S1, T3	S2, T2	S2, T4
U-2	S1, T2	S1, T4	S2, T3	S2, T1
U-3	S1, T3	S1, T1	S2, T4	S2, T2
U-4	S1, T4	S1, T2	S2, T1	S2, T3
U-5	S2,T4	S2, T3	S1, T2	S1, T1
U-6	S2, T1	S2, T4	S1, T3	S1, T2
U-7	S2, T2	S2, T1	S1, T4	S1, T3
U-8	S2, T3	S2, T2	S1, T1	S1, T4

6.2.7 Experimental Procedure

1. An introductory orientation session which asked subjects to read the introduction to the experiment provided on an 'Information Sheet' (Appendix A). This set of instructions was developed to ensure that each subject received precisely the same information.

The first experiment for the explicit feedback system:

2. Subjects filled a pre-search questionnaire, which captured background information on the subject's education, previous general search experience, computer use experience and video Search and general search experience.

3. A training session on the experimental systems with which the subject is to interact, followed by a training topic, which was the same for all subjects. The training session is a chance for subjects to familiarize themselves with the interface components of the experimental systems.

4. Subjects were asked to read the hand-out of written description for the first task (depending on the Greco-Latin design).

5. A session in which the participants interact with the system (depending on the experimental design) in pursuit of the search task they perform. They were given 20 minutes to search and could stop early if they thought that they were unable to find any more relevant information.

6. After completing the search, participants were asked to complete a post-search questionnaire (Appendix F).

7. Participants repeated the steps 4-6 four times.

8. At the end of the experiment, the subject was asked to complete the Final questionnaire (Appendix G).

6.2.8 Training

Participants were asked to do pre-search training because they were unfamiliar to the both experimental systems. About 20 minutes was allocated for training before the start of the experiment. The procedure of the training session went as follows:

1. An introduction to the purpose of the experimental systems.

2. An introduction to the main search interface components that appeared in all experimental systems. Printed screenshots of all experimental systems were used to describe these interface components.

3. A demonstration of each system using the same training search query.

4. The training task gave subjects a chance to familiarise themselves with the main interface components and using the system.

5. The training session ended when subjects felt comfortable using the experimental systems

Subjects had the opportunity to ask questions or comments at any point during the training session. 30 minutes was the maximum time afforded to each subject.

6.2.9 Questionnaire

Questionnaires were the main method used to elicit subject opinion during the experiment. The questionnaires were divided up into the following three sub-questionnaires.

6.2.9.1 Pre-search Question

Through this questionnaire, information about subjects' experience with computers and familiarity with using video was obtained.

6.2.9.2 Post-search Questionnaire

After each task on one of the system given a particular task, the users were asked to complete a questionnaire about the task they were given, the search they performed the system they used, etc.

6.2.9.3 Final Questionnaire

After all experiments, the participants are asked to rank these systems in order of preference with respects to

- The one that helped more in the execution of their tasks,
- The one they liked best.

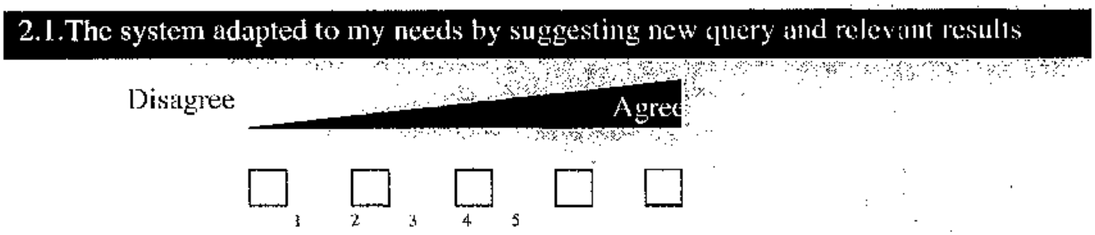
Further, the participants had the opportunity to provide comments about the system being evaluated.

All questionnaires contained three styles of question; *Likert scales*, *semantic differentials* and *open-ended questions*. In this section each style is explained and examples provided. The three sub-questionnaires were divided into a series of subsections that contained questions on the same aspect of the search (e.g., 'Search Process', 'Interface Support'). To help the subject complete the questions, some introductory text was given at the start of each section.

6.2.9.4 Likert Scale

For the purpose of quantifying the expression of agreement or disagreement, I use the five-point likert scale technique, which presents a set of attitude statements. A numerical value from one to five is used to measure each degree of agreement. The attitude can be measured by calculating total numerical value from all responses received.

Figure 6.1 shows an example of Likert scale:



2.1.The system adapted to my needs by suggesting new query and relevant results

Disagree Agre

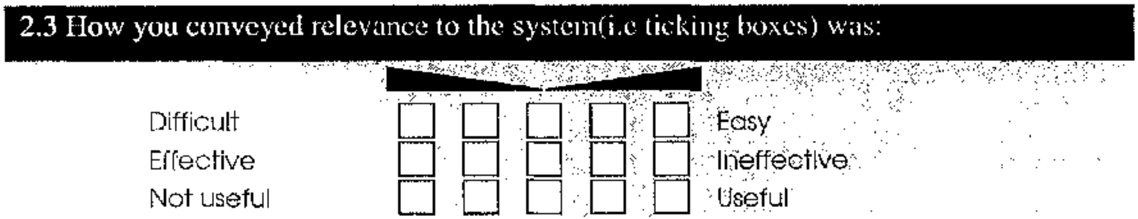
□ □ □ □ □

1 2 3 4 5

6.2.9.5 Semantic Differentials

Semantic Differentials is another type of structured question, which provides pairs of antonyms, together with a five-step rating scales. A pair of words is an object which can express subjects' attitudes. Facing this kind of question, it is a must for subjects to check one of the positions on each continuum between the most positive and negative terms.

Figure 6.2 exemplifies a set of semantic differentials.



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

Difficult □ □ □ □ □ Easy
Effective □ □ □ □ □ Ineffective
Not useful □ □ □ □ □ Useful

6.2.9.6 Open-ended Questions

The style of open-ended questions gives subjects the chance to freely reply without having to select one of several provided options. They are useful for revealing reasons why subjects feel the way they do and giving them a chance to comment freely on various aspects of the system, the task or the experiment generally. An 'Information Sheet' at the start of the search showed subjects completed examples of Likert scales and semantic differentials. It was assumed that subjects could answer unstructured questions without any more instructions on. During the experiment, system logging recorded search activity at the interfaces to the experimental systems. In the next section I describe the logging procedure used.

6.2.9.7 System Logging

When a user is running a search topic, system will automatically log user's actions and related information generated by systems.Log files were named based on the subject's unique identifier, the system and task attempted. The log file is based on XML format, which log the following information:

1. Start and end time of running a search topic;
2. Query type, terms and weights, features type used;
3. The number of terms which are added to user's new query;
4. The number of terms which are removed from user's old query;
5. The number of images which are added to user's new query;
6. The number of images which are removed from user's old query;
7. User's each action which results in a process of updating query.

6.3 Results and Conclusion

This section summarizes the results of the user experiment described in the above sections. The experiment tests two interactive video retrieval systems that have the similar interface but different feedback model support. Experimental subjects attempted search scenarios on the experimental systems and provided feedback on their experience through questionnaires and comments made during informal discussions.

The hypotheses introduced in this chapter are tested in terms of search effectiveness and subject preference. A total of 24 subjects, with different levels of search experience participated in the experiment. The significance of experimental results is tested at $p < .05$ for all tests used, unless otherwise stated. *S1* and *S2* are used to denote the experimental video retrieval systems based with explicit feature and the combination of explicit and implicit features respectively.

The results presented in this chapter are based on questionnaire responses and system logs generated during interaction. The evidence is supported by informal subject feedback and my own observations. Questionnaires used five point Likert scales and semantic differentials with a lower score representing more agreement with the attitude object. The arrangement of positive (e.g., 'easy', 'relaxing') and negative (e.g., 'difficult', 'stressful') descriptors was randomised so that a positive assessment would be represented sometimes by a high score (i.e., approaching 5) and sometimes by a low one (i.e., approaching 1). This ensured that subjects applied due care and attention when completing the differentials (Busha and Harter, 1980). At the analysis stage the high positive scores are reversed so that in all cases the positive assessments were represented by low scores.

No assumptions are made about the normality of the data gathered during the experiment. Non-parametric statistical tests, which are more appropriate than their parametric equivalents, are used to test for statistical significance since these tests do not make any assumptions about the underlying distribution of the data and much of the data gathered was ordinal in nature (e.g., Likert scales and semantic differentials).

I begin this sections by presenting subject demographic and search experience, and results on the search process (Section 6.3.2) and the tasks attempted (Section 6.3.5). Section 6.3.3 presents the results of system, including system effectiveness and relevance assessment. The effectiveness of the way and interface of suggesting terms and images is presented in Section 6.3.4. The results of user's system preference are presented in Section 6.3.6.

6.3.1 Subject Demographics and Search Experience

The average age of the subjects was 27.375 years (maximum 36, minimum 21, standard deviation = 3.76 years). All subjects had a university diploma or a higher degree and were pursuing a qualification in a discipline related to Computing Science. All subjects had rich computing and search experience. All were familiar with web searching service and video search services, and view and watch online news frequently. That shows that all subjects were interested in news and videos and would do the experiment with serious attitudes. Table 6.2 shows the information of all subjects and search experience.

Table 6.2 Subjects characteristics

Factor	Score
Number of Subjects	24
Frequency of Dealing with videos	'once or twice a week' 4.08
Frequency of taking videos	'more than once or twice a month, less than once or twice a week' 3.46
Frequency of carrying out videos searches	'more than once or twice a month, less than once or twice a week' 3.5
Frequency of viewing news	4.21

Frequency of viewing online news	3.79

Table 6.3 (1) Subjects video search experience

Search Engine Used	Google(22),Yahoo(12),AltaVista (3),AlltheWeb(1),Others(9)
Easy/Difficult	3.52
Relaxing/Stressful	3.62
Simple/Simple	3.57
Satisfying/Frustrating	3.19

Note: The number in parentheses is the number of subjects who select the search engine used.

Table 6.3 (2) Subjects video search experience

Question 12. Find your information needs	3.32

Subjects were asked to complete Likert scales asking how much experience they had with video search, such as Google, Yahoo. These results are reported in the table 6.3. Complete semantic differentials on how 'easy'/'difficult', 'stressful'/'relaxing', 'simple'/'complex' and 'satisfying'/'frustrating' the general use of those video search engines were used. The Likert scale values are in the range 1 to 5, where a higher value corresponds to more experience. This was potentially a good indicator of experience levels as I would expect subjects with more experience to be more competent searchers. Table 6.3 (1) showed the average differential responses. Table 6.3 (2) showed the average score for the Question 12.

The popular video search engines for the purpose of video search were satisfied to subjects.

6.3.2 Search Process

In this section I present results on the search subjects performed. Whilst this analysis is not necessary to test the hypotheses, the factors may have an impact on subject perceptions. Each subject was asked to describe various aspects of their experience on each experimental system. The results presented are from questionnaire and informal subject comments, both during the search and after the experiment. Subjects were asked about their search of the information retrieved by each of the experimental systems.

Perceptions of Search

Subjects were asked to complete six semantic differentials about their search: 'relaxing'/'stressful', 'interesting'/'boring', 'restful'/'tiring' 'easy'/'difficult'. 'simple'/'difficult', and 'pleasant /'unpleasant'. The average value in relation to each positive differential is shown in Table 6.4. The 'Overall' value is derived from all six differentials and shows how the process is perceived across all subjects.

Table 6.4 Subject perceptions of the search process (range 1-5, higher = better)

	S1	S2
Relaxing/ stressful	3.71	**3.95**
Interesting/boring	3.88	**4.17**
Restful/ tiring	3.46	**3.67**
Easy/difficult	4	**4.21**
Simple/difficult	**3.88**	3.8
Pleasant/ unpleasant	**4.08**	4

A Friedman Rank Sum Test was run for each differential within all subjects. The test tries to test the difference between the two systems from the perspective of search process.

The results showed no significant differences for all the differentials. Table 6.5 show all $\chi^2(1)$ and the level of significance of the differentials

Table 6.5 $\chi^2(1)$ and the level of significance of the differentials(range 1-5, higher = better).

	$\chi^2(1)$	The level of significance(p)
Relaxing/ stressful	3.27	0.071
Interesting/ boring	1.6	0.206
Restful/ tiring	1.923	0.166
Easy/ difficult	0.692	0.405
Simple/difficult	0.067	0.796
Pleasant/unpleasant	0.111	0.739

The results (Table 6.5) revealed that there is no significant difference between the S1 and S2 from the perspective of search process. All subjects felt the search processing relaxing, interesting, restful, easy, simple, and pleasant in both systems.

6.3.3 System

In this section, I provide the results about system. In order to find difference from the perspective of system, all subjects were asked to answer four questions. Question 2.1 focuses on the effect of adopting user information needs. Question 2.2 is for the relationship between information of a video shot viewed and relevance of the video shot. Question 2.3 provides five differentials for the purpose of measuring the retrieved set. Question 2.4 is used to measure the usefulness of interface. Question 2.5, 2.6, 2.7 and 2.8 are used to compare the two systems from the perspective of the relevance assessment.

6.3.3.1 System Perceived Effectiveness

Table 6.6 shows the average scores of the first two questions of S1 and S2.

Table 6.6 System performance (range 1-5, higher = better).

Question No	S1	S2
2.1	**3.888158**	3.554825
2.2	**3.881944**	3.585648

A Friedman Rank Sum Test was used to test the significant difference between S1 and S2 from user information needs and relationship between viewed information of a video shot and relevance of the video shot. The results show that there is no significant difference between the system based on explicit feature and the system based on the combination of explicit and implicit features (2.1: $X^2(1) = 3.27$, $p = 0.071$; 2.2: $X^2(1) = 3.27$, $p = 0.071$).

Table 6.7 Results of Retrieved Set

	S1	S2
relevant	3.708	**4.125**
important	3.875	**4.125**
useful	4	**4.208**
appropriate	3.667	**3.792**
complete	3.125	**3.917**

With the use of Friedman Rank Sum Test, the results suggested the existence of significant differences on the 'relevant' and 'Complete' differentials (relevant: $X^2(1) = 4.45 > 3.84$, $p =.035 <0.05$, complete: $X^2(1) = 12.25 > 3.84$, $p = 0.0005 < 0.05$), but no difference on the 'important', 'useful', and 'appropriate' differentials (important: $X^2(1) =3.57$, $p =.059$, useful: $X^2(1) = 1.67$, $p = 0.2$, appropriate: $X^2(1) = 0.4$, $p = 0.527$), The results show that the combination system provides more relevant information to subjects and got a significant higher satisfactory degree of completing search topics. But in other aspects, there is no significant difference.

6.3.3.2 Relevance Assessment

The experimental systems differ in how subjects could communicate which information presented at the interface was relevant. The explicit system presents checkboxes, which only allows subjects to explicitly mark relevant items. The combination system is the one which combines implicit assessments of relevance into the explicit system. Subjects were asked about how they told the system which implicit factors (e.g., mouse move over a video shot, viewing textual summary and key-frame of it, playing the video shot) were relevant. Unlike traditional video retrieval systems, it is not a must for subjects to mark a video shot as relevant. The combination system will automatically catch the relevance of video shots according to user's behaviours. The checkbox is also provided because explicit way of marking relevance may allow them to make more accurate relevance assessments. They were asked to complete two kinds of semantic differentials about:

- The *effectiveness* of the assessment method i.e., *How you conveyed relevance to the system was*: 'easy'/'difficult', 'effective'/'ineffective', 'useful'/'not useful'.

- How subjects *felt* about the assessment method i.e., *How you conveyed relevance to the system made you feel:* 'comfortable'/'uncomfortable', 'in control'/'not in control'.

The average obtained differential values are shown in Table 6.8, 6.9 for all subjects.

Table 6.8 Average differential value of conveying relevance to the system (range 1-5, higher = better)

	S1	S2
Easy	4.33	4.04

Effective	**3.83**	3.54
Useful	**4.13**	3.71

A Friedman Rank Sum Test was applied within all subjects, the results of which show that there is no significance between the 'easy' and 'effective' differentials of S1 and S2 (Easy: $X^2(1)$ =1.33, p =.248, Effective: $X^2(1)$ = 1.47, p = 0.225). But the difference of 'useful' differentials between the two systems is significant (Useful: $X^2(1) = 4.0 > 3.84$, $p = 0.046 < 0.05$). This analysis shows that the explicitly marking relevance is significantly more useful than the combination system because of negative effect of the noise which implicit feedback results in. From the perspectives of easiness and effectiveness, there is no significant difference between S1 and S2.

Table 6.9 average differential value of feeling with the relevance convey (range 1-5, higher = better)

	S1	S2
Comfortable	3.83	**4.30**
In control	**3.91**	3.63

There is no significant difference between the two differentials of S1 and S2 by using Friedman Rank Sum Test (Comfortable: $X^2(1)$ =3.77, p =.052, In control: $X^2(1)$ = 0.6, p = 0.439). But the results of Wilcoxon Signed Ranks Test show that the difference of 'comfortable' is significant (Z = -2.217, $p = .027 < 0.05$), the difference of 'in control' is not significant (Z = -1.128, p =.259). The results show that subjects felt comfortable when using S1 and S2. Because of combining the explicit and implicit features, the combination system is also in control. Although the there is no significant difference for the "In Control" differential, most subjects think they can control S1 better than S2.

6.3.4 Additional words and images chosen/recommended by the system

The two systems provide novel interfaces for suggesting user terms and image samples, which automatically update and show the suggested terms and image samples in real time. This section is used to measure the effectiveness and usefulness of the interface and the way in which systems recommend additional terms and images and the difference of these two systems.

They were asked to complete seven questions:

- How subjects *felt* about the suggested terms and images—'I felt comfortable with the way in which the new query was constructed': 'Disagree'/'Agree'(Question3.1);

- The usefulness and relevance of suggested terms and images—e.g., The suggested terms' for query expansion was useful and relevant: 'Disagree'/'Agree'(Question3.2);

- The degree of trusting suggested terms and images—e.g., I would trust the system to choose additional words and images for new search query: 'Disagree'/'Agree'(Question 3.4);

- Comfortable degree of subject with the suggested query—e.g., Felt comfortable with expanded query: 'Disagree'/'Agree' (Question 3.5);

- Subject's idea about the usefulness of interface of recommending search terms and images—I felt the interface of recommending search terms and images was useful: 'Disagree'/'Agree'(Question 3.6);

- Three differentials of the system communication: 'obtrusive'/'unobtrusive', 'informative'/'uninformative', 'untimely'/'timely'(Question 3.7);

- And comments subject can leave.

The average obtained differential values are shown in Table 6.10 and 6.11 for all subjects.

Table 6.10 Average Attitudes to Suggested Terms and Images (range 1-5, higher = better)

Question		S1	S2
3.1	way	3.67	**3.79**
3.2	terms	3.33	**3.71**
3.3	images	3.79	**4**
3.4	trust	3.38	**3.67**
3.5	query	3.67	**3.96**
3.6	interface	4.125	**4.375**

Friedman Rank Sum Tests were used to test the difference of subjects' attitude to the way and interface to suggest terms and images and suggested terms and image samples. The results showed significant differences for the attitude to suggested terms($X^2(1)$ =6.4, p =.011 < 0.05), but no difference for the other questions (way:$X^2(1)$ =.077, p =.782, images:$X^2(1)$ =1.14, p =.285, trust:$X^2(1)$ =1.92, p =.166, query: $X^2(1)$ =1.92, p =.166, interface:$X^2(1)$ =1.33, p =.248). Because there is no value less than 3.0, the attitudes to the way and interface to suggest terms and images, and 'whether subjects trust the suggested terms and images', and 'if subjects felt expanded query comfortable' are positive. The results of Friedman Rank Sum Tests show that subjects' trust to the additional terms suggested by the combination system are much more than subjects' trust to the terms suggested by the system based on explicit feature.

Table 6.11 System Communication (range 1-5, higher = better)

	S1	S2
unobtrusive	2.79	**3.1**
informative	4	**4.17**
timely	**3.54**	3.29

I use the Friedman Rank Sum Tests to test the significant difference between S1 and S2. From the above table, although the average score of the 'unobtrusive' and 'informative' differentials of S2 are higher than that of S1, and the average score of the timely differential of S1 is higher than that of S2, the results of the tests suggested that there is no significant difference between S1 and S2 for the three differentials(unobtrusive:$X^2(1)$ =.286, p =.593, informative: $X^2(1)$ =.692, p =.405, timely:$X^2(1)$ =1.92, p =.166). However, White suggested that implicit feedback systems are unobtrusive and make inferences of what is relevant based on searcher interaction (White *et al.*, 2005). Why is there no significant difference for the 'unobtrusive' differential? One possible reason is that the S2 is a combination system which combines both the explicit and implicit features.

6.3.5 Task

In this section the search tasks, attempted by experimental subjects, was discussed. Tasks were divided into four categories proposed by Yan (Rong, Yang, and Hauptmann *et al.*, 2004) and within these categories into four search topics. In order to encourage naturalistic search behaviour, simulated situations are appropriate. Simulated situations proposed in (Borlund, 2000a; Borlund, 2000b), which can reflect and simulate a real information seeking situation, were applied for this purpose of putting simple search topics into a real situation. Figure 6.3 shows an example simulated situation.

Figure 6.3 Simulated Situation

Simulated Situation

Assume that you are a tour guide of the Tomb of the Unknown Soldier at Arlington National Cemetery. You are going to give an introduction of the Tomb of the Unknown Soldier at Arlington National Cemetery to visitors before visiting the cemetery. Please find as many video shots of the Tomb of the Unknown Soldier at Arlington National Cemetery as possible to make your presentation.

Table 6.12 Differentials of Search Tasks (range 1-5, higher = better)

Question		S1	S2
4.1	Clear	4.29	**4.56**
	Simple	4.13	**4.25**
	Familiar	4	4
4.2	More information	2.8	**3.17**
4.3	Difficulty	4.04	**4.08**
4.4	Success	3.67	**4.29**

Table 6.13 Significant Difference Tests (range 1-5, higher = better)

Question		$X^2(1)$	Level of significance (p-value)
4.1	Clear	3.27	0.071
	Simple	1.33	0.248
	Familiar	1	1
4.2	more information	1.92	0.166
4.3	Difficulty	0.818	0.366
4.4	Believe	4	**0.046 < 0.05**

Table 6.12 and 6.13 shows the results of Friedman Rank Sum Tests for the question 4.1-4.4. The results of the Friedman Rank Sum Tests suggested that the significant differences for the success in the performance of the search tasks attempted by experimental subjects between S1 and S2. This means that the participants have much more belief that they have succeeded in their performance of this task. The results suggested that there is no significant difference for the three differentials--'clear', 'simple', and 'familiar' between S1 and S2. It also suggested that there is no significant difference for the difficulty of the search tasks which were running on S1 and S2. This means that these search tasks are clear, simple, familiar to subjects, and have similar difficulty.

6.3.6 System Preference

In this section, I analyze the results of the exit questionnaire/interview. The analysis based on quantitative data and Qualitative Data was proposed in section 6.3.6.1 and 6.3.6.2 respectively.

6.3.6.1 Quantitative Data

Subjects used each of the two systems and were asked to rank them in their order of preference without any instructions were given when subjects were making their decision. Subjects were asked to give a brief explanation of their ordering. In Table 6.14, the rank order of the two systems (S1, S2) is presented for all subjects.

Table 6.14 Rank order of systems (range 1-2, lower = better)

	S1	S2
Rank	1.83	**1.17**

Kruskal-Wallis test, Wilcoxon Signed Ranks Test, and Friedman Test were used to test the significant difference for the ranking between the two systems. The results of all these three statistical tests suggested that there is a significant difference for the ranking of these two systems. Table 6.15 shows the results of three statistical tests.

Table 6.15 Results of Three Statistical Tests

Statistical Test	Value	Level of Significance (p-value)
Friedman Test	**$X^2(1) = 9.78$**	**$0.00176 < 0.05$**
Wilcoxon Signed Ranks Test	**$Z = -3.13$**	**$0.00176 < 0.05$**
Kruskal Wallis Test	**$X^2(1) = 22$**	**2.73E-06 < 0.05**

6.3.6.2 Qualitative Data

All subjects were asked to provide a brief explanation about their ranking and leave some comments about their experience on these two systems.

Though the analysis of quantitative data revealed most of subjects preferred the system based on the combination of explicit and implicit features. However, there are 4 subjects who preferred the explicit system, and one subject was not sure which system is better. In this section, I briefly introduce the reasons of subjects' ranking from two perspectives.

The main notion of participants who prefer the system based on explicit feature (S1) is that they like the system which they can control very well and felt the first system was good at being in control, although statistical tests show that there is no significant difference between these two systems for the differential. One subject felt the automatic way of the combination system to capture subjects' behaviors "confused". The one who is not sure which system is better has the idea that the performance of a video retrieval system is dependent on the search tasks. The result of Koenemann and Belkin *et al.*, (1996) also showed that people prefer control.

Most of the participants think the second system is better than the first one because they prefer the automatic way of the combination system to capture subjects' behaviors and like the automatic suggestions from the system since they suggested that the automatic manner make the search easy, and it is not necessary for subjects to do much relevance assessment explicitly by themselves.

6.3.7 System Logging

In this section, the analysis of system log files will be presented.

Table 16 shows the results of information logged when all users were using the two systems.

Table 6.16 System log

	S1	S2
Sum of Added Terms	2384	**4362**
Sum of Removed Terms	1191	**3084**
Sum of Added Images	293	**458**
Sum of Removed Images	87	**226**
Average number of Added Terms	12.749	**29.275**
Average number of Removed Terms	6.37	**20.7**
Average number of Added images	1.57	**3.07**
Average number of Removed Image	0.465	**1.52**
Average Iterations/each topic	4.02	**3.27**
Sum of Iterations	193	**157**
Average Duration of each task (ms)	455.2768	**329.369**

The result shows that more terms and images were automatically added to user's new query by S2. S2 added 4362 terms (29.275/per iteration), and 458 images (3.07/per iteration) to user's new query totally. More terms were removed from user's old query. S1 removed average 6.37terms per iteration from user's old query, but S2 removed 20.7 terms per iteration from user's old query. That means that the frequency of updating query terms of S2 is really higher than S1.

The average duration of performing each task of S2 is shorter than S1. The average duration of performing each task of S2 is only 329.369. The result of Wilcoxon Signed Ranks test shows a significant difference (Z =-3.66159, Asymp. Sig. (2-tailed) = 0.000251 < 0.05) between S1 and S2 on the average duration of performing each task. Friedman Test also shows the significant difference (Chi-Square = 10.08333333, N = 48, df = 1, Asymp. Sig.= 0.001496164 < 0.05)between S1 and S2. When user uses S1, average number of iterations for each search task is 4.02, but when user uses S2 , average number of iterations for each search task is 3.27. Both Friedman test (Z=-2.092767936, Asymp. Sig. (2-tailed) = 0.036369875) and Wilcoxon Signed Ranks test (N = 48, Chi-Square = 4, df = 1, Asymp. Sig. = 0.045500264) show that there is significant difference for the number of iterations for each search topic between S1 and S2. The analysis of system log files show S2 has better performance than S1.

6.4 Summary

In this chapter I have presented and analysed the findings of the user experiment. The user experiment aimed to compare the effectiveness of the two interactive video retrieval systems, the usefulness of the interface for a video retrieval system.

The complete results of the user experiment reflected that there is significant difference for only six differentials. This means that the combination system (S2) is a little bit better than the explicit system only in some aspects. The first experimental hypothesis at the beginning of this chapter can be supported in part by this user experiment.

The other two experimental hypotheses were also supported by the analysis of results. The system preference of almost all subjects is S2. This means that subjects preferred the combination video retrieval system. The novel interface for suggesting terms and image samples are also useful and effective for subjects. It is also proven that the user's actions (e.g. playing a video shot, browsing video, or seeing related information of one video shot) in a video retrieval system can be considered as useful evidence for the relevance of video shots and obvious indicators which can reflect users' interests during a search session.

Chapter 7

Conclusion and Future work

7.1 Introduction

In this thesis I have investigated the use of implicit feedback techniques, which help searchers to create new queries and effectively use these new queries to find new video shots, to help searchers use interactive video retrieval systems more effectively. In Chapter 3, I described an experiment to benchmark the role of various features for video retrieval. In Chapter 5, I introduced the heuristic-based implicit feedback models which capture searchers' behaviour during a search session and evaluated the models based on simulation-based methodology. Chapter 6 presented the results of the user experiment, which is used to evaluate the effectiveness of the interface of suggesting terms and images, and compare the usefulness of the interactive video retrieval system based on the combination of explicit and implicit features and the explicit-based video retrieval system. In this chapter, I make a conclusion and summary of the main findings and contributions of this thesis and future work.

7.2 Query Categorization

In Chapter 3, I used two query categorization schemes to benchmark the effectiveness of various features for video retrieval.

In order to use effectively appropriate features for video retrieval systems, the investigation on the relationship between query categories and features were designed. Two sets of experiments were conducted based on these two query categorization schemes. Results of these system experiments reveal the potential relationship between query categories and features and the consistency between these two approaches. It is that using the text-based feature or both text-based features and image features really outperform the systems using image features only. It is necessary for a video retrieval system to use text-based features. Enough text-based features can make sure a higher performance than that of the system using image features only. For the specific categories (e.g., Specific Object Category which Yan defines) the performance of the system using both kinds of features is higher; and, the performance of the system using both image features and text-based features is same as the system based on text-based features in the Scene and Event category only. The conclusion from this study is that considering the cost for computing similarities based on image features, which are described as a multi-dimensions vector, its role in video retrieval is prominently useful only in some special query categories. It is beneficial to detect such categories automatically and employ specific retrieval strategies.

7.3 Implicit factors and Implicit Feedback Models

In Chapter 5, I proposed four implicit factors in the context of a video retrieval system. Those four implicit factors are derived from possible behaviour when a searcher is using a video retrieval system. Based on these four implicit factors, some heuristic based implicit feedback models were proposed. All implicit feedback models evaluated in the study increased search effectiveness through query modification. The Binary Voting Model performs particularly well. Furthermore, from the aspect of marginal effects, pure Binary Voting Model also has the largest marginal effect. The implicit feedback models make it possible for searchers to automatically expand query, when a search has changed based on short-term, within search session, interaction data.

A simulation-based evaluation methodology was used to benchmark the performance of implicit feedback models. Four implicit feedback models were tested totally by using the methodology. This methodology has the advantages of less time consuming and costly compared with user experiment, the requirements of which are more strict and complex. It is easy to model searcher's interaction with video retrieval systems and test the performance of a number of implicit feedback models and find the best performance model, which will be deployed in the experimentation with human subjects. But the effectiveness of system interfaces can not be evaluated by this strategy.

For the purpose of evaluating implicit feedback models from user's perspective, a user experiment involved 24 subjects was designed. Its result reflected that the heuristic-based implicit feedback models that choose new query terms and image samples for query expansion are useful and appropriate. Some of the problems inherent in traditional RF could be mitigated by the techniques discussed in this thesis. For example, searchers are directly involved in the explicit relevance assessment. It is possible that the initial query is automatically modified for satisfying a searcher's need based on an iterative process of feedback without explicit relevance assessment. The next section discusses the effectiveness of simulation-based evaluation methodology.

The analysis of results based on the user experiment shows that interface is useful for users to do video search. The second system based on the combination of explicit and implicit features is the preference of most subjects. Both of these systems provided a Checkbox system that relied on explicit relevance assessments. The reason of providing explicit feature in an implicit system is that the explicit relevance assessments are more relevant. Implicit feedback has more possibility of obtaining noise information by capturing user's interactions than the explicit system. The interface for suggesting terms and images for a video retrieval system has been proven to be useful. Suggested terms were showed in a text editable area. Users are allowed to add or remove any suggested terms. Suggested images were showed in a scrollable panel. Each image was visualized by a thumbnail with a checkbox, which was designed for users to re-construct an image-based query by removing any images.

7.4 Future Work

In this section, I discuss possible future work to improve the search effectiveness of video retrieval systems, usefulness of interfaces from the following three perspectives.

From the perspective of implicit indicators, I only investigated four implicit factors, which are the most essential behaviour in a context of video retrieval. In a video retrieval context, there are more implicit potential factors which indicate the relevance of video shots. For example, forward play, backward play, slow play, adjusting volume of sound stream of a video file, adjusting the properties of a frame (e.g. colour, luminance, contrast and so on). Differences between various video players make it possible to have different kinds of operations on a video file. The way of computing a weight for an implicit factor in this thesis is the simplest method. There should be some sophisticated methods that can be used to compute a weight for an implicit factor. Serious investigations are needed to find such methods.

The query categorization is based on the search topics defined by TRECVID2003. It is a specific collection of search topics. Is this categorization also effective for more general search topics? It is certain that there must be more possibly useful query categorization which should be investigated deeply. In addition, techniques need to be developed to find such categories automatically.

From the perspective of interfaces, there are many aspects which can be improved. For example, the two systems do not have progress bar for showing the search progress. It is possible that there is much better way of suggesting terms and images, and the time when system suggests terms and images. Frequency and the form of suggesting are also very important issues which should be tested in further user experiments.

Reference

Ardizzone , E. , Cascia, M. La and Molinelli, D. , (1996), Motion and color based video indexing and retrieval. In *Int. Conf. on Pattern Recognition (ICPR'96)*, Vienna, Austria.

Arthur M., S.D., Brodley, C.E., Shyu, C.R., (2000), Relevance feedback decision trees in content-based image retrieval, *Proc. of IEEE Workshop on Content-based Access of Image and Video Libraries*,(pp.68-72).

Bates, M., (1989), The design of browsing and berry-picking techniques for the online search interface. *Online Review*, 13 (5), (pp.407-424).

Belkin, N. J. and Vickery, A. (1985). *Interaction in information system: A review of research from document retrieval to knowledge-based system*, 188-198. London: The British Library.

Beaulieu, M. and Walker, S., (1992), An evaluation of automatic query expansion in an online library catalog. *Journal of Documentation*, 48, (pp.406-421).

Beaulieu, M. (1997), Experiments on interfaces to support query expansion. *Journal of Documentation*, 53 (1), (pp.8-19).

Belkin, N. J., Cool, C., Koenemann, J., Bor Ng, K. and Park, S. Y., (1996b), Using relevance feedback and ranking in interactive searching. *Proceedings of the Fourth Text Retrieval Conference*. (pp.181-210).

Belkin, N. J., (2000), Helping people find what they don't know. *Communications of the ACM*, 43 (8), (pp.59-61).

Belkin, N. J., Cool, C., Kelly, D., Lin, S.-J., Park, S.-Y., Perez-Carballo, J., (2001). Iterative exploration, design and evaluation for query reformulation in interactive information retrieval. *Information Processing and Management*, 37, (pp.403-434).

Budzik, J., and Hammond, K., (1999), Watson: Anticipating and contextualizing information needs. In *Proceedings of the 62nd Meeting of the American Society for Information Science*, USA, (pp.727-740).

Borlund, P. and Ingwersen, P., (1997), The development of a method for the evaluation of interactive information retrieval systems. In *Journal of Documentation*, Vol. 53, no. 3, (pp.225-250).

Borlund, P. (2000a), *Evaluation of interactive information retrieval systems*. Unpublished doctoral dissertation, Åbo Akademi University,

Borlund, P., (2000b), Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 56 (1), (pp. 71-90).

Browne, P., Czirjek, C., Gaughan, G., Gurrin, C., Jones, G. J.F., Lee, H., Marlow, S., McDonald, K., Murphy, N., O'Connor, N.E., O'Hare, N., Smeaton, A.F., Ye, J. M., (2003), *Dublin City University Video Track Experiments for TREC 2003*

Buckley, C., Salton, G. and Allan, J., (1994), The effect of adding relevance information in a relevance feedback environment. *Proceedings of the 17th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp.92-300).

Busha, C. H. and Harter, S. P. (1980) *Research methods in librarianship: Techniques and interpretation*. New York: Academic Press Inc.

Campbell, I. and Van Rijsbergen, C. J., (1996), The ostensive model of developing information needs. *Proceedings of the 3rd International Conference on Conceptions of Library and Information Science*. (pp.251-268).

Canon, (2003), *Local Sales of Still Digital Cameras Jump by 2000%* (2003), Retrieved December 30, 2003 from http://www.canon.com.au/home/story_893.html.

Carbonell , J., Geng, G., and Goldstein, J. (1997), Automated query- Relevant Summarization and Diversity-Digital-Based Reranking. In *IJCAI-97 Workshop on AI and Digital Libraries*

Chi, E. H., Pirolli, P., Chen, K. and Pitkow, J., (2001), Using Information Scent to Model User Information Needs and Actions on the Web. *Proceedings of the Conference on Human Factors in Computer Systems*, (pp.490-497).

Chi, E. H., Rosien, A., Supattanasiri, G., Williams, A., Royer, C., Chow, C., (2003), The Bloodhound Project: Automating Discovery of Web Usability Issues using the InfoScent Simulator. *Proceedings of the Conference on Human Factors in Computer Systems*, (pp.505-512)

Claypool, M., Le, P., Waseda, M. and Brown, D., (2001), Implicit interest indicators. *Proceedings of the 6th International Conference on Intelligent User Interfaces*. (pp.33-40).

Cleverdon, C. W. (1960) *Aslib Cranfield research project: Report on the first stage of an investigation into the comparative efficiency of indexing systems*. Cranfield: The College of Aeronautics.

Cleverdon, C.W., Mills, J. and Keen, M., (1966), *Factors Determining the Performance of Indexing Systems*, Volume I - Design, Volume II - Test Results, ASLIB Cranfield Project, Cranfield.

Efron, B. and Morris, C., (1977), Stein's parador in statistics, Scientific American, 236, (pp.119-127)

Egan, D., Remde, J., Gomez, L., Landaur, T., Eberhardt, J. and Lochbaum, C., (1989), Formative Design-Evaluation of SuperBook. *ACM Transactions on Information Systems*, 7:1, (pp.30-57).

Fagin, R., Kumar, R., and Sivakumar, D.., (2003), Efficient similarity search and classification via rank aggregation. In *Proceedings Of the ACM SIGMOD Int. Conf. on Management of Data*, (pp.301–312).

Feng, D., Siu, W.C., Zhang, H.J., (2003), *Multimedia Information Retrieval and Management-Technological Fundamental and Application*, Springer-Verlag Berlin Heidelberg, New York

Fuhr,N., (1992), Probabilistic models in information retrieval. *The Computer Journal*, 35(3), (pp.243-255).

Giorgio, B., Stefano, M., and Carlo T., (1996), "Evaluating User Interfaces to Information Retrieval Systems: A Case Study on User Support". In *SIGIR96, 19th International Conference on Research and Development in Information Retrieval*, pp128-136, Zurich, Switzerland, (pp.18-22).

Golovchinsky, G., Price, M. N., and Schilit, B. N., (1999), From reading to retrieval: Freeform ink annotations as queries. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*, USA, (pp.19-25).

Grossman, D.A., Froeder, O., (1998), *INFORMATION RETRIEVAL Algorithms and Heuristics*, Boston/Dordrecht/London, Kluwer Academic Publishers.

Gruss, R., (2004), *The VisOR System: Testing the utility of user interface components for feature-based searching in video retrieval software.* A Master's paper for the M.S. in I.S. degree. School of Information and Library Science, University of North Carolina at Chapel Hill. January, 2004.

Harman, D. (1993). Overview of the first TREC conference. *Proceedings of the 16th Annual ACM SIGIR Conference of Research and Development in Information Retrieval.* (pp. 36-47).

Harman, D., (1988), Towards interactive query expansion. *Proceedings of the 11th Annual ACM SIGIR Conference on Research and Development in Information Retrieval.* (pp.321-331).

Harper, D. J., (1980), *Relevance Feedback in Document Retrieval Systems.* Unpublished doctoral dissertation, University of Cambridge, Cambridge, UK.

Hauptmann, A., Baron, R.V., Chen, M.Y., Christel, M., Duygulu P., Huang, C., R., Lin, W.H., Ng, T., Moraveji, N., Papernick, N., Snoek, C.G.M., Zanetakis, G., Yang, J., Yang, R., and Wactlar, H.D., (2003), *Informedia at TRECVID 2003: Analyzing and Searching Broadcast News Video*

Hersh, W. R., Elliot, D.L., Hickam, D.H., Wolf, S. L., and Molnar, A., (1995), Towards new measures of information retrieval evaluation. In *Proceedings of the 18th Annual International ACM/SIGIR Conference*, (pp.164-170), Seattle, WA

Jose, J. M. (1998), *An Integrated Approach for Multimedia Information Retrieval.* PhD thesis, The Robert Gordon University, Aberdeen, Apr..

Jones, W., Bruce, H. and Dumais, S. (2001). Keeping found things found on the web. *Proceedings of the 10th Conference on Information and Knowledge Management.* (pp.119-134).

Kang, I.-H. and Kim, G., (2003), Query type classification for web document retrieval. *In Proc. of the 26th ACM SIGIR*, (pp.64-71). ACM Press.

Kelly, D. and Teevan, J., (2003), Implicit feedback for inferring user preference. *SIGIR Forum*, 37 (2), (pp.18-28).

Kelly, D., (2004), *Understanding implicit feedback and document preference: A naturalistic user study.* Unpublished doctoral dissertation, Rutgers University, New Jersey.

Kleinberg, J. M., (1999), Authoritative sources in a hyperlinked environment.
Journal of the ACM, 46(5), (pp.604-632).

Key-frame, http://www.clickandgovideo.ac.uk/Glossary.htm

Koenemann, J. and Belkin, N. J., (1996), A case for interaction: A study of interactive information retrieval behavior and effectiveness. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems.* (pp.205-212).

Lee H., (2001), *User Interface Design for Key frame-Based Content Browsing of Digital Video*, Ph.D. Thesis, Dublin City University

Li, X. and Roth, D., (2002), Learning question classifiers. In *COLING'02*.

LIMSI, (Gauvain, J.L., Lamel, L., and Adda, G..), (2002), The LIMSI Broadcast News Transcription System. *Speech Communication*, 37(1-2):89-108, ftp://tlp.limsi.fr/public/spcH4_limsi.ps.Z

Lancaster, F.W., (1996), Evaluation of interactive knowledge-based systems: Overview and design for empirical testing. *Journal of the American Society for Information Science*, 47(1), (pp.57-69).

Magennis, M. and van Rijsbergen, C. J., (1998), The potential and actual effectiveness of interactive query expansion. *Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 324-332).

Marchionini G. and Geisler G., (2002), *The Open Video Digital Library*, D-Lib Magazine, Vol. 8, Number 12, December.

Maron, M. and Kuhns, J., (1960), On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing machines*, 7: (pp.216-244).

Morita, M. and Shinoda, Y., (1994), Information filtering based on user behavior analysis and best match text retrieval. *Proceedings of the 17th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 272-281).

Mostafa, J., Mukhopadhyay, S. and Palakal, M., (2003), Simulation studies of different dimensions of users' interests and their impact on user modelling and information filtering. *Information Retrieval*, 6, (pp.199-223).

MySQL Internals Manual 4.7 Full-text Search http://dev.mysql.com/doc/internals/en/full-text-search.html

Nichols, D. M., (1997), Implicit ratings and filtering. *Proceedings of the 5th DELOS Workshop on Filtering and Collaborative Filtering*. (pp. 31-36).

Oard, D. and Kim, J., (2001), Modeling information content using observable behaviors. *Proceedings of the 64th Annual Meeting of the American Society for Information Science and Technology.* (pp. 38-45).

Pirolli, P. and Card, S. (1995), Information foraging in information access environments. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems.* (pp. 51-58).

Porkaew K., Chakrabarti K. and Mehrotra S., (1999), Query refinement for multimedia similarity retrieval in MARS. In *Proc. of the ACM Int. Conf. on Multimedia,* (pp.235–238), Orlando, Florida.

ReplayTV. Available online at URL: http://www.replaytv.com(last visited August 2005)

Rijsbergen, C.J.V., (1979), *Information Retrieval,* London: Butterworths, 1979.

Rijsbergen, C. J. Van, (1986), A new theoretical framework for information retrieval. *Proceedings of the 10th International ACM SIGIR Conference on Research and Development in Information Retrieval.* (pp.194-200).

Rocchio, J.J. (1971), Relevance feedback in information retrieval in The SMART Retrieval system Experiments in Automatic Document Processing, Prentice Hall (pp.313-323)

Robertson, S.E. and Theater, D., (1974), A statistical analysis of retrieval tests: a Bayesian approach, *Journal of Documentation,* **30,** (pp.273-282)

Robertson. S. E., (1986), On relevance weight estimation and query expansion. *Journal of Documentation*, 42, (pp.182-188).

Robertson, S.E., Walker, S., and Jones, M.S., (1995), Okapi at TREC-3, *Proceedings of Second Text Retrieval Conf.* (TREC-3)

Rong Y., Yang, J., Hauptmann, A.G., (2004), Learning query-class dependent weights in automatic video retrieval, *Proceedings of the 12th annual ACM international conference on Multimedia*, October 10-16, New York, NY, USA

Ruthven, I., (2001), *Abduction, explanation and relevance feedback.* Unpublished doctoral dissertation, University of Glasgow, Glasgow, UK.

Ruthven, I., Lalmas, M. and Van Rijsbergen, C. J., (2002b), Ranking expansion terms using partial and ostensive relevance. *Proceedings of the 4th International Conference on Conceptions of Library and Information Science.* (pp. 199-219).

Ruthven, I., (2003), Re-examining the potential effectiveness of interactive query expansion. *Proceedings of the 26th ACM SIGIR Conference on Research and Development in Information Retrieval.* (pp. 213-220).

Ruthven, I., Lalmas, M. and Van Rijisbengen, C. J., (2003), Incorporating user search behaviour into relevance feedback. *Journal of the American Society for Information Science and Technology*, 54 (6), (pp.528-548).

Salton, G. Yang, C., and Wong, A., (1975),. A vector-space model for automatic indexing . *Communications of the ACM*, 18(11), (pp.613-620).

Salton, G., (1989), *Automatic Text Processing*, Addison-Wesley.

Salton, G. and Buckley, C. (1990), Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science.* 41. 4. (pp. 288-297).

Shen, X.H., and Zhai C.X., (2003), Active Feedback – *UIUC TREC-2003 HARD Experiments*, University of Illinois at Urbana-Champaign,

Singhal, A. and Buckley, C. and Mitra, M., (1996), Pivoted Document Length Normalization, *ACM SIGIR'96*, (pp.21-29).

Spärck-Jones, K. (1981). Retrieval system tests 1958-1978. (Ed, Spärck-Jones, K.), *Information retrieval experiment* (pp. 213-255). London: Butterworths.

Spink, A., Goodrum, A., & Hurson, A. R. (2001). Multimedia we queries: Implications for design. *Proceedings of the International Conference of Information Technology: Coding and Computing*, Las Vegas, NV, (pp.589 – 593).

Sonka, M., Hlavac, V., and Boyle, R.. (1998), *Image Processing, Analysis, and Machine Vision*. Brooks and Cole Publishing, 2nd edition.

Stéphane Marchand-Maillet, (2002), *Content-based Video Retrieval: An overview*, Viper team CUI - Université de Genève.

Stricker, M., Orengo, M.,(1995), Similarity of color images. In *Proc. of the SPIE: Storage and Retrieval for Image and Video Databases*. Volume 2420. (pp. 381-392).

Swanson, D. R. (1986). Subjective versus objective relevance in bibliographic retrieval systems. *Library Quarterly*, 56, (pp.389-398).

Tague, S.J. and Nelson M.J., (1981), Simulation of User Judgements in Bibliographic Retrieval Systems, *Proceedings of the 4th annual international ACM SIGIR conference on Information storage and retrieval: theoretical issues in information retrieval*, (pp.66-71)

Tague S. J., (1992), The pragmatics of information retrieval experimentation, revisited. *Information Processing and Management*, 28 (4), (pp.467-490).

TiVo, Available online at URL: http://www.tivo.com (last visited Augest 2005)
TRECVID, (2003) http://www-nlpir.nist.gov/projects/tv2003/tv2003.html

Urban J. and Jose J.M., (2004), M., Evidence Combination for Multi-Point Query Learning in Content-Based Image *In Proceedings of the IEEE Sixth Int. Symposium on Multimedia Software Engineering*, (pp. 583--586)

Wactlar, H., Christe. M., Gong, Y., Hauptmann A. (1999), Lessons Learned from Building a Terabyte Digital Video Library. *IEEE Computer, Special Issue on Digital Libraries*, 32(2), pp. 66-63.

White, R.W., Jose, J.M. and Ruthven, I., (2002), The use of implicit evidence for relevance feedback in Web retrieval. *Proceedings of 24th ECIR Conference*, 93-109..

White, R. W., Ruthven, I. and Jose, J. M., (2002b), The use of implicit evidence for relevance feedback in web retrieval. *Proceedings of 24th BCS-IRSG European Colloquium on Information Retrieval Research*. (pp. 93-109).

White, R.W., Jose, J.M, van Rijisbergen, C.J. and Ruthven, I., (2003), A Simulated Study of Implicit Feedback Models, *16th European Conference on IR Research, ECIR 2004*, Sunderland, UK, April 2004 Proceedings, Springer, (pp.311-326).

White, R. W. and Jose, J. M., (2004), A study of topic similarity measures. *Proceedings of the 27thAnnual ACM Conference on Research and Development in Information Retrieval*. (pp. 520-521).

White, R. W., (2005), *Implicit Feedback for Interactive Information Retrieval*, Ph.D. Thesis. University Of Glasgow

Wildemuth, B. M., Yang, M., Hughes, A., Gruss, R., Geisler, G., and Marchionini, G., (2003), Access via Features versus Access via Transcripts: User Performance and Satisfaction, *TREC VID 2003 Notebook Paper SILS Technical Report*

Xu, J., (1997), Solving the Word Mismatch Problem Through Automatic Text Analysis, PhD Thesis, University of Massachusetts at Amherst.

Yang, H., Chaisorn, L., Zhao, Y., Neo, S.Y., and Chua, T.S.., (2003), VideoQA: question answering on news video. In *Proc. of the 11th ACM MM*, (pp.632-641).

Zhang, H.J., (1995), "Video parsing, retrieval and browsing: an integrated and content-based solution," Proc. ACM Multimedia'95, San Francisco. Nov. 5-9, (pp.15-24).

Zhang, D. and Lee, W. S., (2003), Question classification using support vector machines. *In Proc. of the 26th ACM SIGIR*, (pp 26-32), ACM Press

Zhong, Y., (2000), *Apply Multimodal Search and Relevance Feedback in a Digital Video Library*, A Thesis Master of Science in Information Networking.

Appendix A—Information Sheet

Title of Project:

Interactive video retrieval system

UNIVERSITY
of
GLASGOW

Name of Researcher:

Huang, Zheng

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

The aim of this experiment is to investigate the relative effectiveness of two different video retrieval 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 video retrieval systems and their interface, not you, that are being evaluated. You were chosen, along with 24 others, because you work or study at the University of Glasgow. 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. If you decide not to take part 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. A decision not to participate will not affect your grades in any way.

The experiment will last about one hour and will you will receive a reward of £10 upon completion. You will be given a chance to learn how to use the two interfaces before we begin. At this time you will also be asked to complete an introductory questionnaire. You will perform four tasks, one with each system, and complete a questionnaire about using each system. The questionnaires will ask how you felt during each search. All of your interaction (e.g., mouse clicks, playing, mouse movement) will also be logged. You are encouraged to comment on each interface as you use it, all your comments will be recorded by taking notes *if you so prefer*. You will have the option to review, edit, or erase the recording. Please ask questions if you need to and please let me know when you are finished each task. You will be asked some questions about the tasks and systems at the end of the experiment.

All information which is collected about you during the course of this research 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 only for analysis, then destroyed. The results of this study will be used for my Mrc research. The results are likely to be published in late 2005. 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 research is being funded by the Research Committee at the Department of Computing Science, University of Glasgow. This project has been reviewed by the Faulty of Information and Mathematical Sciences Ethics Committee. For further information about this experiment please contact:

Huang, Zheng(e.mail: hzheng@dcs.gla.ac.uk or tel: 0141 330 2788).
Department of Computing Science, University of Glasgow
17 Lilybank Gardens
Glasgow, G12 8RZ.
27/07/05Information Sheet (Version 1.0)

Appendix B—Consent Form

CONSENT FORM

UNIVERSITY
of
GLASGOW

Title of Project:

Interactive Video retrieval System

Name of Researcher:
Huang Zheng

1. I confirm I have read and understand the information sheet dated □

(..../..../2005) (version ....) for the above study and have had the opportunity to ask questions.

2. I understand that my permission is voluntary and that I am free to withdraw at any time, without giving any reason, without my legal rights being affected. □

3. I agree to take part in the above study. □

4. I would like to receive a summary sheet of the experimental findings □

IF YOU WISH A SUMMARY, leave an email address:______________________________
Name of subject Date Signature:______________
Researcher Date Signature:________________
1 for subject; 1 for researcher
Please initial box

Department: *Computing Science*

Subject Identification Number for this study:______________

Appendix C—Receipt of Payment

Department: *Computing Science*
Subject Identification Number for this study:

RECEIPT OF PAYMENT

UNIVERSITY
of
GLASGOW

Title of Project:

EVALUATION OF THE INTERACTIVE VIDEO RETRIEVAL SYSTEMS BASED ON EXPLICIT FEATURE AND THE CO AND IMPLICIT FEATURES

Name of Researcher:

Huang, Zheng

I confirm receipt of £10 paid for my participation in the above experiment.

Name of subject   Date   Signature

Researcher   Date   Signature

Appendix D—Task Description

TASK A

0103-- Find shots of Yasser Arafat

Task Description:

Assume that you are a journalist working for a local newspaper. You are writing an article on the role of Yasser Arafat in Israle Palestinian Conflict after he died this year. Find as many relevant video shots of Yasser Arafat as possible to complete your article.

Background:

Yasser Arafat is the most famous leader of Palestine Liberation Organization. Palestine Liberation Organization was founded in 1964 as a Palestinian nationalist umbrella organization dedicated to the establishment of an independent Palestinian state. After the 1967 Arab-Israeli war, control devolved to the leadership of the various fedayeen militia groups, the most dominant of which was Yasser Arafat's Al-Fatah. In 1969, Arafat became chairman of the PLO's Executive Committee. The Israeli-Palestinian conflict, a part of the greater Arab-Israeli conflict, is an ongoing conflict between Israel and Palestinians. The Israeli-Palestinian conflict is by no means a simple two-sided conflict with all Israelis (or even all Israeli Jews) sharing one point of view and all Palestinians another. In both communities, there are individuals and groups who advocate total territorial removal of the other community, those who advocate a two-state solution, and those who advocate a binational solution of a single secular state encompassing present-day Israel and the Gaza strip and the West Bank.

TASK B

0109-- Find shots of one or more tanks

Task Description:

Assume you are going to give a talk at the local veterans group. You are going to talk about the role of 'tanks' in modern warfare. Find as many video shots of tanks as possible.

Background:

A tank is a tracked, armoured combat vehicle (armoured fighting vehicle), designed primarily to destroy enemy ground forces by direct fire. A modern main battle tank (MBT) is distinguished by its high level of firepower, mobility and armour protection relative to other vehicles of an era. It has the heaviest armour of any vehicle on the battlefield, and carries what is intended to be an effective anti-tank weapon. It is among the most versatile and fearsome weapons on the battlefield, valued for its shock action against other troops, its ability to engage a wide variety of ground targets, and high survivability. Tanks can be vulnerable if not properly protected from other weapons especially aircraft strikes, mines, and artillery, as well as being swamped by infantry. They are usually employed as part of combined arms warfare, supported by infantry, other fighting vehicles and aircraft.

Tanks were first used in World War I, to break the deadlock of the trenches, and they evolved to take the role of cavalry on the battlefield. The name "tank" first arose in British factories making the casings of the first battle tanks: the workmen were given the impression they were constructing tracked water containers for the British Army, hence keeping the production of a fighting vehicle in secret. Tanks have subsequently undergone many generations of design evolution; many of their traits have matured. However, there is an ongoing arms race between tank armour and anti-tank weapons systems, and between opposing tank designs, causing a continual need for upgrading.

TASK C

0106-- Find shots of the Tomb of the Unknown Soldier at Arlington National Cemetery

Task description:

Assume that you are a tour guide of the Tomb of the Unknown Soldier at Arlington National Cemetery. You are going to give an introduction of the Tomb of the Unknown Soldier at Arlington National Cemetery to visitors before visiting the cemetery. Please find as many video shots of the Tomb of the Unknown Soldier at Arlington National Cemetery as possible to make your presentation.

Background:

The Tomb of the Unknowns, near the center of the cemetery, is one of Arlington's most popular tourist sites. The Tomb contains the remains of unknown American soldiers from World Wars I and II, the Korean Conflict and (until 1998) the Vietnam War. Each was presented with the Medal of Honor at the time of interment and the medals, as well as the flags which covered their caskets, are on display inside the Memorial Amphitheater, directly to the rear of the Tomb. The Tomb is guarded 24-hours-per-day and 365-days-per year by specially trained members of the 3rd United States Infantry (The Old Guard). The Memorial Amphitheater has been the scene of the funerals of some prominent Americans (such as General John J. "Black Jack" Pershing) as well as the site of both Memorial Day and Veterans Days celebrations.

TASK D

0117-- Find shots of one or more groups of people, a crowd, walking in an urban environment (for example with streets, traffic, and/or buildings)

Task Description:

Assume that you are a researcher for a magazine about urban life. You are assisting your editor who is writing an article about environment. Please find as many relevant video shots of groups of people, street, traffic or buildings in an urban environment to help your editor to write the article.

Appendix E—Entry Questionnaire

ENTRY QUESTIONNAIRE

UNIVERSIT
of
GLASGOW

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.. □ Female... □

3. Please provide your current OCCUPATION: YEAR :

4. What is your FIELD of work or study?

Part 2: SEARCH EXPERIENCE

Experience with Videos

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
5. deal with videos in your work, study or spare time?	1	2	3	4	5	6
6. take videos in your work, study or spare time?	1	2	3	4	5	6
7. carry out video searches at home or work?	1	2	3	4	5	6

Video Search Experience

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

None.. □

Google (http://www.google.com).. □

Yahoo (http://www.yahoo.com).. □

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

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

Others (Please specify) □

9. How often do you view news

Never

□ □ □ □ □

1 2 3 4 5

N/A □

10. How often do you watch/view online news

Never

□ □ □ □ □

1 2 3 4 5

N/A □

11. Using the video search services you chose in question 8 is GENERALLY:

easy	□	□	□	□	□	Difficult
stressful	□	□	□	□	□	Relaxing
simple	□	□	□	□	□	Complex
satisfying	□	□	□	□	□	Frustrating

N/A □

12.You find what you are searching for on any kind of video search service…

Never

					N/A
□	□	□	□	□	□
1	2	3	4	5	

Appendix F—Post-search Questionnaire

Post-Search Questionnaire

UNIVERSIT
of
GLASGOW

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.

Section 1: Search Process

1.1 I felt this search process was:

Relaxing	□	□	□	□	□	Stressful
Interesting	□	□	□	□	□	Boring
Restful	□	□	□	□	□	Tiring
Easy	□	□	□	□	□	Difficult
Simple	□	□	□	□	□	Complex
Pleasant	□	□	□	□	□	Unpleasant

Section 2: System

Each system of two systems has different features to help you find relevant information. In this section I ask you about the system you have just used.

2.1 The system adapted to my needs by suggesting new query and relevant results

Disagree Agree

□ □ □ □ □

1 2 3 4 5

2.2 The more information you viewed from a video shot, the more relevant the video shot is:

Disagree Agree

□ □ □ □ □

1 2 3 4 5

2.3 I felt the retrieved set was:

Relevant	□	□	□	□	□	Irrelevant
Important	□	□	□	□	□	Unimportant
Useful	□	□	□	□	□	Useless
Appropriate	□	□	□	□	□	Inappropriate
Complete	□	□	□	□	□	Incomplete

2.4 The interface layout was useful:

Disagree Agree

□	□	□	□	□
1	2	3	4	5

Comments:

Relevance Assessment:

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

Difficult	□	□	□	□	□	Easy
Effective	□	□	□	□	□	Ineffective
Not useful	□	□	□	□	□	Useful

2.6 How you conveyed relevance to the system was

Difficult	□	□	□	□	□	Simple
Useful	□	□	□	□	□	Not useful

2.7 How you convey relevance to the system made you feel

Comfortable	□	□	□	□	□	Uncomfortable
Not in control	□	□	□	□	□	In control

Comment n your experience with system features? (For combination system)

2.8 I am happy with the way system informed me of new query

Disagree Agree

□ □ □ □ □

1 2 3 4 5

Section 3: Additional words and images chosen/recommended by the system

The system chose or recommended additional query words and image sample query. In this section I ask you about this process

3.1 I felt comfortable with the way in which the new query was constructed:

Disagree Agree

□ □ □ □ □

1 2 3 4 5

3.2 The suggested terms for query expansion was useful and relevant

Disagree Agree

□	□	□	□	□
1	2	3	4	5

3.3 The suggested images for query expansion was useful and relevant

Disagree Agree

□	□	□	□	□
1	2	3	4	5

3.4 I would trust the system to choose additional words and images for new search query:

Disagree Agree

□	□	□	□	□
1	2	3	4	5

3.5 Felt comfortable with the expanded query:

Disagree Agree

□	□	□	□	□
1	2	3	4	5

3.6 I felt the interface of recommending search terms and images was useful:

Disagree Agree

□	□	□	□	□
1	2	3	4	5

3.7 The system communication is in a way that was:

obtrusive	□	□	□	□	□	unobtrusive
informative	□	□	□	□	□	uninformative
untimely	□	□	□	□	□	timely

3.8 Do you have any further comments about the words and images chosen/recommended?

Section 4: Task:

In this section I ask about the search task you have just attempted.

4.1 The search task was:

Unclear	□	□	□	□	□	Clear
Simple	□	□	□	□	□	Complex
Unfamiliar	□	□	□	□	□	Familiar

4.2 I think there was better information available (that the system did not help me find)

Disagree Agree

□ □ □ □ □

1 2 3 4 5

4.3 I would rate the difficulty of the tasks

Easy □ □ □ □ □ Difficult

4.4 I believe I have succeeded in my performance of this task

Disagree Agree

□ □ □ □ □

1 2 3 4 5

4.5 Do you have any further comments about the tasks you have just attempted?

Appendix G—Exit Questionnaire

Exit Questionnaire/Interview

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

UNIVERSITY
of
GLASGOW

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.

Section 1: System Experiences

1.1. Rank the systems in order of preference(1 = best, 2 = worst)?

System 1 (Explicit feature-based system):

System 2 (the system based on a combination of explicit and implicit features):

1.2. Please comment on the above rankings.

(a) what was the main reason?

(b) Any other comments?

1.3. Comment on your experience with each system?

a. System 1(Explicit feature-based system)

b. System 2(the system based on a combination of explicit and implicit features)

Section 2: Comments

2.1 Do you have any further comments or questions about the systems or experiment

Please take note of my email address and let me know if you have any further questions.

Thank you for your help