Adaptive Systems for Multimedia Information Retrieval

Gareth J. F. Jones

School of Computing, Dublin City University
Dublin 9, Ireland
Gareth.Jones@computing.dcu.ie

Abstract. Multimedia information retrieval poses both technical and design challenges beyond those of established text retrieval. These issues extend both to the entry of search requests, system interation and the browsing of retrieved content, and the methodologies and techniques for content indexing. Prototype multimedia information retrieval systems are currently being developed which enable the exploration of both the user interaction and technical issues. The suitability of the solutions developed within these systems is currently being explored in the annual TRECVID evaluation workshops which enable researchers to test their indexing and retrieval algorithms and complete systems on common tasks and datasets.

1 Introduction

The rapid expansion in the availability of online multimedia content has led to a similarly rapid growth in research into technologies for automated retrieval of multimedia information. The potential for exciting new multimedia applications targeted at operational environments ranges from entertainment and education to academic research and intelligence services. The possibilities of what these systems might achieve is to a significant extent limited only by the imagination of system developers. Systems for multimedia information retrieval are by their nature complex typically requiring the integration and adaptation of a number of existing technologies as well as the development of novel algorithms and techniques. The success of realizing these visions of what might be is limited both by the availability of the required technologies, but also to a considerable degree by the quality of the analysis of user and system requirements and the consequent design of the retrieval application.

The diversity of multimedia content types and the variety of environments for which multimedia information retrieval systems might be developed means that there is no single ideal solution. It is thus vital when exploring the development of a new multimedia information retrieval application to properly understand the user and their need for the application (whether or not this is an existing need or an application created one), the capabilities of the hardware environment in which the application must operate, and the available retrieval and information management technologies.

In order to better understand the importance of all these components of a multimedia information retrieval system this paper first explores the possible definitions of multimedia information retrieval and the importance of adaption in these systems, then briefly examines some important issues relating to both multimedia systems and user tasks, considers issues relating to retrieval and content indexing, then goes on to demonstrate the application of some of these features within the Fischlár system being developed in the Centre for Digital Video Processing (CDVP) at Dublin CIty University, the next section then outlines the international TRECVID task for evaluating and understanding current video retrieval technologies, the paper ends with some concluding thoughts on future research directions.

2 What is a Multimedia Information Retrieval System?

One important question is, what constitutes a multimedia information retrieval system? It is often assumed that it must involve retrieval of full motion video, but it is sometimes referred to in the context of the retrieval of spoken documents without reference to associated video content or the retrieval of static images. In relation to video and image retrieval, multimedia retrieval process itself often involves only the analysis of linguistic material associated with the visual content, in the case of video the spoken soundtrack and the use of textual labels for images. Where visual data is present, it is natural to think in terms of analysis of this content and its use in the retrieval process. However, as we will see later this is much less straightforward than might at first be assumed.

Another significant issue in respect of the definition of multimedia retrieval is to consider what the system must actually be capable of doing. Established text information retrieval systems, as exemplified by web search engines, use a user search request to compute a ranked list of potentially relevant documents which is returned to the user with a short piece of text from each document that hopefully reliable indicates the main reason why this document has been adjudged to be potentially relevant. The user then selects the document that they feel is most likey to satisfy their information need, and downloads the document to read it and extract the necessary information from it. It may seem obvious, but in the context of extending this paradigm to multimedia data is important to appreciate that, the user currently addresses their information need by reading the whole text document, with perhaps a small amount of keyword searching if it is a long document. One possibility for a multimedia information retrieval system is merely to replicate this text searching environment requiring the user to audition retrieved documents to find the relevant information. However, several features of multimedia data mean that these systems must be both more complex and include the user in a much more integrated way.

Firstly, in the case of temporal media such as video and spoken data, browsing large amounts of material is very time consuming since the user must listen to it. Playback can be faster than real-time, but even doubling the speed of delivery does not begin to address this problem. Second, when considering visual

media there is the fundamental question of how the user should express their information need. In the case of linguistic data it is natural to assume that a written search request is an appropriate form of expression, although this may not always be the case, for example if the user is uncertain about domain specific terminology. Browsing of temporal multimedia documents is addressed in most video and audio retrieval systems by the development of a graphical browsing application which aims to direct users to potentially relevant sections of the document without having to play back the document in its entirety from the beginning. These browsers typically show time on a horizontal graphical bar representing a complete retrieval unit, e.g. a document, with potential points of interest marked along the bar [1] [2]. Further indication of the content can be given in the case of video content by the use of a series of keyframes taken from the video which are shown in a single screen. Typically clicking any selected point on the document bar will begin play back from that position in the document.

How though should the user express a need for information contained in non-linguistic form for visual content searching? Perhaps they can express their request in text which is then matched against automatically generated labels of the visual content. Or perhaps they can sketch what they are looking for and an image search performed. Or if they happen to have an existing image example this could be used in a "query by example" framework to find similar images. These approaches make various assumptions about the sophistication of automatic content indexing, the user's ability to express what they are looking for or the availability of existing exemplars of what they are looking for. As we will see later feature extraction is one of the most significant challenges facing visual based multimedia information retrieval, and probably one of the reasons that many such systems make heavy reliance on the use of linguistic content. One of the weaknesses of this dependence on linguistic content is that spoken soundtracks and textual image labels will in general only express a very limited interpretation of the visual content, and in some cases will bear no relationship to the visual content at all. Thus there is a very real need for the development of visual indexing technologies.

The difficulty in expressing information needs for visual content and the limitations in visual indexing mean that the searching process will often need to be much more interactive with the user involved in multiple cycles of query refinement to actually find what they are looking for. For image retrieval this will typically involve a combination of searching on textual labels and then refinement by selecting images that are related to the desired image using query by example feedback cycles. For video retrieval a similar process will be carried out using content from the spoken soundtrack and feedback using keyframes from the video and potentially complete scenes. The limitation and unreliablity of video feature extraction and the usual importance of any associated linguistic content means that retrieval decisions should generally be based on a combination of matching scores derived from multiple media streams.

The complexity of the retrieval and browsing phases of multimedia information retrieval makes it attractive to make use of any additional information that

might be available. Thus the system should make use of explicit user feedback from the current search: e.g. "this document is relevant", "this person may be important in relevant documents"; implicit feedback, e.g. playback of an entire document often suggests that it is relevant or at least partially relevant; or the user's previous searching history. Feedback methods are considered further in Section 5.1.

3 Multimedia Systems

The broadest definition of a multimedia system usually involves the potential to deliver visual and audio content to a user as required. This may mean delivery from a local source such as a DVD or CD-ROM, or playback across a network which may itself only be a local area network or a much larger wide area network. The fidelity of the content that can be delivered will depend on both the computational resources available at each point in the network and also the bandwidth of the network itself.

Until fairly recently a multimedia system would involve a high-power computer connected to a hard-wired network. However, this situation is rapidly evolving to include broadband wireless networks and the capability of multimedia processing on handheld computing devices such as PDAs and mobile telephones. The various networking technologies involved in connecting these devices to the network have different bandwidths and latency specifications, the computing devices themselves have varying resources for data processing and differing physical resources for information delivery, and importantly the users of these different platforms are working in a variety of different environments.

All these issues taken together mean that multimedia information retrieval applications must be appropriate to the network, the hardware being used and the user's physical environment. Thus the applications should adapt to the multimedia system being used. For example, the fidelity of the content delivery should not exceed the capacity of the network or the computing device, and the user interface to the system should take into account the physical dimensions of the computing device enabling the user to view the output easily on small devices while not restricting the possibility for complex interaction and visualization on desktop systems.

4 User Tasks

The specifications of the multimedia platforms and networks, and the available indexing and retrieval technologies only provides the potential to develop effective multimedia information retrieval applications. It is vital in attempting to specify useful applications that developers analyze the needs and potential needs of the users of these applications.

It is often argued in respect of user interface design for computing applications that these should be based on a careful analysis of the tasks that users will really wish to carry out, and that this should include concepts and vocabulary

with which the target user group are already familiar. This is often referred to as adopting a strategy of *user centered design*.

While this is certainly true of multimedia information retrieval systems, since users will often not be familiar with applications of the type that we are trying to develop, it seems inevitable that new concepts will be introduced that users will not be familiar with. In this case it is important that these novel concepts are ones which build on those with which the user is already familiar. It is often tempting to develop applications for developers, and not those targeted at real users. I suspect that this is particularly true of multimedia information retrieval applications and I would caution developers to always bear this point in mind. For example, while it may seem attractive to develop interfaces which adapt automatically using machine learning techniques based on input from user behaviour, these modified interfaces are unlikely to find favour with users if they cannot work out how to perform operations because basic interface consistency principles are being broken in the adaptation process. Much guidance on these issues is available on the user interface design literature [3].

5 Information Retrieval

A full description of information retrieval methods is beyond the scope of this paper, this section highlights some relevant features from text retrieval methods that can be applicable for multimedia applications.

Text retrieval systems are usually based on computing a matching score between some form of textual search request and each available document in an archive. A list of documents ranked by matching score is then returned to the user. There are a number of elaborations on this approach are available to improve performance or adapt the method to different tasks. These include relevance feedback methods, personalization, and recommender systems and collaborative filtering.

5.1 Relevance Feedback

Relevance feedback methods provide a number of possible techniques to adapt user search requests. The input to the relevance feedback process is the existing search request, the set of documents returned in response to this query, and the user's judgements of the relevance of these retrieved documents. The output is typically a ranked set of possible expansion terms that may be added to the existing request and information to modify some of the parameters of the search system to enhance the ranking of documents similar to those marked relevant in the current search. It is hoped that adding the proposed additional terms to the request will make it a better expression of the user information need, and that modifying the search parameters will promote the rank of further relevant documents. The basic underlying assumption being that further relevant documents will in some way resemble those already identified.

The search request can be expanded automatically to include the highest ranked of the proposed terms or the terms can be offered to the user for them to select the terms which they feel best reflect concepts related to their information need [4]. The expanded query statement is then applied to the search archive and a new ranked list retrieved. The dominant effect in relevance feedback usually relates to query expansion, but its effectiveness is usually enhanced by its combination with modification of search term weights to favour terms associated with relevant documents.

An alternative to interactive relevance feedback to *pseudo* relevance feedback where a number of the top ranked documents in the initial retrieval pass are assumed to be relevant, the expansion terms and revised search term weights are computed as before but assuming the relevant document set, and then performing another retrieval pass with this revised topic statement before presenting the revised ranked list to the user. Of course, some of the documents assumed to be relevant will not in fact be relevant, this can lead to selection of some poor expansion terms which can actually reduce performance for the second retrieval pass. On average the effect is generally observed to be beneficial to retrieval accuracy, but it can be disastrous for individual queries particularly if none of the assumed documents are in fact relevant. True relevance feedback based on users' relevance judgements will in general be better.

5.2 Personalization

Relevance feedback usually refers only to the adaptation of retrieval systems parameters and the request for a single ad hoc request. A more elaborate use of feedback information is to provide a personalization of the retrieval system to the individual user. Where this is done the retrieval system will adapt to the behaviour of individual users possibly over a single searching session or over an extended period of time, or a combination of both, and its response to a search request will be different for each user.

The basic process of personalization is to use previous relevance judgements to develop one or more profiles associated with each user that represents their ongoing interests, e.g. particular sports teams or news topics. Profiles are typically a set of keywords which may be weighted based on their perceived importance in expressing the user's interests. A variety of methods are possible to form these profiles and utilize them in searching.

Personalization Agents One approach to personalization of retrieval systems is to make use of agents to model user interests. One example of such a system is *Amalthaea* developed at the MIT Media Laboratory [5]. This system is based on an ecosystem of evolving agents which represent user interests. Agents are rewarded for delivering relevant documents to the user and the best agents reproduce by using the genetic methods of mutation and crossover. The lowest scoring agents are purged from the system with the aim of maintaining a gene

pool of consistent size which best models the user's current interests. Experimental studies show that Amalthaea is able to rapidly adapt to changes in the user's interests.

5.3 Recommender Systems and Collaborative Filtering

Personalization based on the behaviour of individual users relies on the limited amount of information that can be gathered based on their actions. An alternative is to gather information from a number of equivalent users and combine this information to represent their shared interests. These group profiles are thus based on a broad based set of user experiences and in general more relevance data. This data can be used to recommend potentially relevant material and also within interactive retrieval [6].

Individual and group profiles can be combined to give personal adaptive profiles with contribution from group experiences.

6 Multimedia Information Retrieval

The previous section introduced some adaptation methods used in text retrieval systems. While these methods are all currently used in many prototype systems, they remain the subject of active research interest. This section looks at existing multimedia information retrieval and considers how adaptation techniques have been applied to date and how they might be further extended.

6.1 Spoken Document Retrieval

The most mature area of multimedia information retrieval relates to spoken documents. If the contents of spoken documents are fully manually transcribed, the retrieval stage would be a standard text retrieval problem. However, manual transcription of more than a trivial amount of spoken content is generally prohibitively expensive (domains such as mass media broadcast TV or film are a notable exception) and spoken document retrieval systems thus usually rely on transcriptions generated by automatic speech recognition systems. Various approaches to speech recognition for indexing spoken documents for retrieval have been explored, but comparative experiments have demonstrated that formation of a full transcription using large vocabulary recognition gives the best output for retrieval purposes [7], at least in the domain explored of TV news. It is not clear whether this is the optimal indexing solution for less structured data with a vocabulary less well matched to the document domain. As with all speech recognition systems, transcription systems make errors in their output. The number of errors is related to the quality and content of the audio signal and typically varies from around 5% to over 80% with an average using current systems of around 20%. It has been found in experiments that this level of errors in the transcription has only a very small impact on retrieval accuracy [7].

For some data sources, such as TV broadcasts, textual closed captions or subtitles of the audio are broadcast with the audio-video material. While often not a perfect transcription of the audio material, the quality is usually better than that generated automatically using speech recognition. The closed-captions can be decoded into standard text and used as the search index data. The closed captioning is usually not closely aligned to the actual audio data. However, a forced alignment speech recognition phase can be used align the audio content to the closed captions enabling fine granularity searching.

Where the number of indexing errors is sufficiently high to impact retrieval performance, it is generally found that pseudo relevance feedback methods are particularly effective for improving the spoken document retrieval [8].

6.2 Image and Video Retrieval

Retrieving multimedia content using information extracted from visual media is much more challenging technically in terms of feature extraction, but also from the perspective of user interaction. Images frequently have many interpretations, some of these can be measured directly from visual features, but often the intended interpretation will depend on the context in which the image is being viewed. For example, an object may be identified as a building, as a cathedral, as a specific named cathedral, as being of a particular style and period of architecture, or by the religious denomination to which it belongs. Some of these interpretations can be made directly from analysis of the image using suitable templates, others would require additional information sources to be consulted. Even to carry out the image only interpretation requires the image to be indexed using appropriate features.

In principle if a standard set of features could be agreed, and it is not at all clear that this could be possible, then all video content could be manually annotated. However, the cost of doing this would be uneconomic for all but the most important data. Thus automated feature indexing is likely to be even more important than for spoken content indexing.

Automatic feature extraction for image and video data is currently the focus of a large research effort, but so far the achievements remain very limited. For image retrieval indexing is often based on extraction of colour histograms with a limited amount of spatial information included. Much research is also exploring specific feature extraction tools, often relating to specific domains, for example identification of named people or people in general, cars, sky, etc. Video analysis often includes structural indexing such as the detection of shot boundaries, and attempting to identify keyframes from within identified shots. The same colour histogram and feature extraction techniques are typically applied to individual frames. Ideally features should be derived automatically, robust, accurate and above all useful for retrieval. It would be nice to build feature detectors for each query as it is entered, but this is not practical and retrieval must make use of the feature analysis carried out when the data was initially indexed. Systems are typically configured to only attempt to recognise the presence of a very limited number of features. The limited number of features and the difficulty

in defining features that are in generic means that image analysis systems are domain specific. This may be a very tightly specified domain, e.g. recognising the presence of a moving care in a video, or broader (but nevertheless limited to a specific task), e.g. retrieving images from a collection of disjoint photographs using matching of colour regions.

The matching of identified visual features and search queries is so far fairly simple by comparison to text retrieval for which a number theoretically motivated models have been developed.

One of the important areas for multimedia information retrieval is the integration of data from the multiple media streams, e.g. audio, visual and textual. Evidence of each media stream can be used to reinforce each other for more accurate feature identification and retrieval. For example, a name individual may be recognised by uncertain evidence from visual recognition, speaker identification from the audio stream, and the name appearing in accompanying textual information. If all these sources indicate the presence of a named individual, there is a very good chance that this person is indeed present in this shot. Research in this area is again at an early stage, but encouraging results are being achieved using techniques such as Support Vector Machines [9].

Moving beyond recognition of objects in narrow domains to identification, tracking and interpretation of objects within complex video is a long term research goal of those working in this area.

Relevance Feedback Difficulties in defining search requests and indexing mean that user feedback in an iterative searching process is an important topic for multimedia information retrieval. Feedback can be used for adaptation in the search, both by applying it to the linguistic data, as is already widely exploited, but also to the image data. An example of relevance feedback for image retrieval is described in [10].

An important observation with respect to relevance assessment is that judging relevance of textual documents takes a small, but potentially significant amount of time, particularly is a large number of documents must be judged. However, user assessment of the relevance of an image is almost instantaneous. Thus, while it is unclear how to specify a search request properly representative of the user information need for visual search, it is comparatively easy to obtain large amounts of relevance data from the user in response to initial search runs.

This suggests that there is a much greater need for the "user-in-the-loop" for searching of visual media. Thus while fully automated searching may not be effective due to the ambiguity in image interpretation, much relevance data can be collected for each search to better understand the user's need in this context [11].

Given that the purpose of feedback here is not so much to "find more like this" as is often assumed to be the case for text relevance feedback, but rather to learn more about what is required in a relevance image, it makes sense to show the user the most-informative images for feedback, which may not coincide with the most-positive images for an individual search. This can be thought

of as the differences between a "show-me-the-results" type display versus an "ask-me-questions" user model [11].

The difficulties of specifying information need and the ease with which users can make relevance judgements of images are two of the motivating factors in the proposed use of the *ostensive models* of relevance feedback for image search as described in [12]. The ostensive model supports a query-less interfaces in which the user's indication of the relevance of the retrieved objects is used as the indication of the user's current information need. Therefore it allows direct searching without the need to formally specify the information need.

Image Clustering It is sometimes argued that images can be clustered to assist browsing of similar images to one identified as relevant. However, semantically meaningful clustering depends on the subspace in which a semantic concept class lies, in the case of images semantically meaningful classes will depend largely on the user's current interpretation of the image relevance. It is this information that relevance feedback is trying to capture, so prior clustering of images may often be of limited utility for image searching.

An illustration of the issue of preclustering is provided by [13]. Images are clustered based on colour features, manual examination of the clusters quickly reveals inappropriate groupings. However, moving images between clusters to correct these mistakes and reclustering incorporating this feedback information leads to more reliable clusters. This provides an example of the importance and effectiveness of involving the human user in the management of image data.

A number of systems are currently in development to explore issues in multimedia information retrieval and to investigate the development of techniques to improve multimedia information retrieval performance. One such system is the Fischlár-News Video being developed within the Centre for Digital Video Processing at Dublin City University (DCU). The next section outlines Fischlár highlighting its current use of adaptation. The next section outlines the current Fischlár prototype system, and the following one introduces the TRECVID international evaluation exercise and discusses the application of Fischlár to this task.

7 Fischlár-News

Fischlár is a digital video archive system based on MPEG-7 digital video content management and retrieval, and supports playback using both fixed and mobile computing devices. Fischlár is deployed across the university campus at DCU and currently has more than 1000 regular users.

Fischlár-News automatically records the 30 minutes 9.00pm news every day from the Irish national broadcaster RTÉ1. The system is accessible on campus via any web browser [14] and is now being made available on mobile devices [15]. Currently several months of recorded news is available online with more than 2 years of material in the overall archive.

Fig. 1. Architecture of Fischlár News.

In order to support access from different platforms Fischlár-News is based on XML technologies, which by incorporating XSL transformations for each new device required, can easily be extended to incorporate new access technologies, devices and standards. Figure 1 shows the basic architecture of Fischlár-News illustrating both desktop and mobile access and the process of automatic news story segmentation.

In Fischlár-News mobile access is supported for both PDAs (Campaq iPAQ on a wireless LAN) and XDAs. Each of these play RealVideo encoded content encoded at 20Kbps in order to support streaming access a mobile phone network on an XDA. The desktop version of the system uses a conventional web browser with MPEG-1 video streaming.

Fischlár currently performs shot boundary detection on the captured data and identifies scenes and keyframes. One area of current work is on the automated segmentation of news broadcasts into story level units which typically combine a number of separate shots [9]. Although work is progressing well on this, the current prototype system relies on manual segmentation of each broadcast into story units.

7.1 Fischlár-News on the Desktop

The basic level of browsing on desktop Fischlár-News is the news broadcast. The basic display is shown in Figure 2. Selecting a broadcast from the calender on the left hand side displays a list of news stories from this broadcast. Each news story within the broadcast is represented by a keyframe and textual description of the story.

When presented with a list of news stories the user has the option to either play back the news story by clicking on the "PLAY THIS STORY" button

Fig. 2. Fischlár News with stories from one program.

(Figure 3) or to examine the story at the shot level by selecting the keyframe or the numbered news story title. This produces a detailed list of camera shots and associated closed caption text (Figure 4).

The size of the broadcast archive means that content-based searching is vital in Fischlár News. Each news story is represented by a textual description automatically extracted from the closed caption text broadcast along with the audio and video data. This textual description is used for story-level searching of the archive. The search query is entered in the query box at the top left hand corner of the interface.

A ranked list of stories is returned and displayed in the story column on the right hand side of the interface. Individual stories can be played back and browsed at the shot level as before. However, in addition in this case the story can be viewed in the context of that day's news broadcast by following the date link which displays a listing of news stories from the news program for that day.

Using the closed caption transcripts similar stories are identified and links formed between the related stories. Fischlár-News generates a ranked list of the ten most similar news stories. These related stories can be shown in a ranked story list on the right hand side of the display and browsed as for the other options.

In order to provide personalization and recommendations, user feedback is collected. At any point while browsing the archive or an individual story the user is presented with the opportunity to rate the story on a five point scale from "do not like" to "like very much". In addition to gathering explicit user feedback, usage data is collected automatically as the user plays back or browses

Fig. 3. Playing back a news story.

news stories. Recommendations from users are used as one of the primary access mechanisms for the mobile version of the Fischlár-News system.

7.2 Fischlár-News on Mobile Devices

The small display size and the difficulties of data input for mobile devices, as well as the observation that users are often engaged in distractive environments while using these devices, present major constraints in the design of mobile applications. In consequence it has been suggested that different interaction paradigms are suitable for mobile devices, rather than just porting the interaction methods from the desktop environment. Based on various user studies some general design principles for interaction with mobile devices have been proposed. These include principles such as minimizing the required user input, e.g. provide yes/no options, hyperlinks, and filtering the available information to deliver only content that is most likely to be relevant to this user.

These guidelines suggest that more pre-processing of the information should be carried out by the system prior to delivery to the user. For example, increasing the use of recommender technologies so that material is delivered with less user interaction. This is particularly important for multimedia information retrieval where browsing is such an important issue in information access. The Fischlár-News mobile application uses the personalized list of news stories as the primary access point for mobile users [9].

The starting point for access on the mobile Fischlár-News system is the personalized story list shown in Figure 5. The only input that is required from the

Fig. 4. Shot-level browsing of a news story.

user is to select a news story to play back using RealVideo, shown in Figure 6. This approach minimizes the user input by filtering out content that this user may not be interested in.

As an alternative to the personalized list, the user can be presented with a reverse chronological listing of recorded news programmes, shown in Figure 7. This enables them to view the entire archive. When a broadcast is selected it is presented to the user as a listing of composite news stories, as shown in Figure 8.

8 TRECVID - Benchmarking Video Retrieval

TREC (Text REtrieval Conference) is an annual research exercise organised by the US NIST which enables groups to perform comparative experiments on information retrieval tasks. TREC culminates in a workshop where participants gather to report their results and the methods used to achieve them. TREC, now in its 12th year, has explored a wide range of information retrieval tasks including ad hoc retrieval, web retrieval, cross-language retrieval, question-answering, interactive retrieval and spoken document retrieval. TREC is a global activity with around 100 groups now participating in one or more of the tasks.

Since 2001 TREC has included the TRECVID track which aims to explore video data retrieval. Each year TRECVID makes an agreed set of data available to registered participants. A number of tasks are then performed on this data. These tasks have included: shot boundary detection, semantic feature extraction, news story segmentation, and interactive searching for relevant video associated

Fig. 5. Personalized story recommendation.

Fig. 6. Playback on a mobile device.

Fig. 7. Reverse chronological daily news listing on a mobile device.

Fig. 8. Story listing on a specific date.

with a set of predefined topics. The organizers and participants develop a set of perfect results or relevant stories, and all submissions are scored relative to these ideal results.

TRECVID has confirmed video retrieval as a very challenging task and participants have worked collaboratively both to markup the data for development and testing, and shared their work. Thus groups have shared the task of marking up data with the features present in each shot and shared the output of their automatic feature extraction tools to enable greater research exploration by the community. Features explored in TRECVID so far include: outdoors/indoors, face detection, people detection, cityscape, landscape, speech, instrument sound. Search topics for the interactive evaluation include issues such as: "locomotive approaching the viewer," and "microscopic views of living cells".

The CDVP at DCU participation in TRECVID 2003 is extending the prototype Fischlár system described in the previous section. The group is participating primarily in two tasks: the automated story boundary detection task using a technique based on video analysis and support vector machines, and the interactive searching task.

For the interactive task a group of test subjects were given a number of the search topics released by the TRECVID organisers at NIST. The test data for TRECVID 2003 was a set of news broadcasts from CNN and ABC. The objective is to locate as many relevant items as possible within a limited search time. Each broadcast was segmented by NIST into a standard set of shots with a keyframe for each shot. In order to support this task the interface outlined in the previous section was extended to include user marking of relevant shots retrieved in response to the initial search query. The text from this shot and information automatically taken from analysis of the keyframe for the shot were used in a relevance feedback adaptation of the initial search to expand the search query. The features taken from the keyframe were based on a combination of regional colour histogram analysis, average and maximum regional colour and regional edges. In the feedback search run the expanded text query was matched against shots and a matching score was also computed between the keyframe from each relevant shot and other keyframes in the collection. The scores generated from the text and images searches were then combined to form an overall score for each shot. A revised retrieved list was then presented to the test user for further relevance judgement and, if needed, further iteration of the query. Full details of the DCU participation in TRECVID 2003 are contained in [16].

9 Concluding Remarks

Multimedia information retrieval is a challenging research area which despite recent progress will continue to present research challenges for many years. The complexity of the searching task both in terms of specification of information need and location of relevant content means that adaptation via relevance feedback, personalization and collaborative filtering is potentially very useful for these tasks. Further, the proliferation of devices with the capability for multi-

media playback and searching means that interfaces and search modalities must be adapted to take account of the differing interactivity constraints of these devices. While it may be possible to do this to some extent automatically this needs to be handled with care in order to maintain interface consistency between different platforms.

Current prototype systems demonstrate that it is already possible to build large scale networked multimedia information retrieval systems. The quality and bandwidth of networked computing is set to increase further in the coming years. There are interesting challenges in terms of developing new multimedia information management tools and interfaces appropriate for the many platforms that are available. By far the most significant technical challenge at this point would appear to lie with the automated extraction and labelling of features and their use in the retrieval process.

Acknowledgements

I am grateful to members of the CDVP at DCU for discussing latest developments in Fischlár, and Cathal Gurrin for supplying the images used in this paper.

References

[1] M. G. Brown, J. T. Foote, G. J. F. Jones, K. Sparck Jones, and S. J. Young. Automatic Content-Based Retrieval of Broadcast News. In *Proceedings of ACM Multimedia 95*, pages 35–43, San Francisco, 1995.

[2] M. G. Brown, J. T. Foote, G. J. F. Jones, K. Sparck Jones, and S. J. Young. Open-vocabulary Speech Indexing for Voice and Video Mail Retrieval. In *Proceedings of ACM Multimedia 96*, pages 307–316, Boston, 1996.

[3] B. Shneiderman. *Designing the User Interface*, Addison-Wesley, 1998.

[4] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In D. K. Harman, editor, *The Third Text REtrieval Conference (TREC-3)*, pages 109–126. NIST, 1995.

[5] A. Moukas and P. Maes. Amalthaea: An Evolving Multiagent Information Filtering and Discovery System for the WWW. *Journal of Autonomous Agents and Multi-Agent Systems*, 1(1):59-88, 1998.

[6] B. Smyth and P. Cotter. A Personalized Television Listings Service. *Communications of the ACM*, 43(8), 2000.

[7] J. S. Garafolo, C. G. P. Auzanne, and E. M.@Voorhees. The TREC Spoken Document RetrievalTrack: A Success Story. In *Proceedings of the RIAO 2000 Conference: Content-Based Multimedia Information Access*, pages 120, Paris, 2000.

[8] G. J. F. Jones and A. M. Lam-Adesina. An Investigation of Mixed-Media Information Retrieval. In *Proceedings of the Sixth European Conference on Research and Advanced Technology for Digital Libraries*, pages 463-478, Rome, 2002.

[9] C. Gurrin, A. F. Smeaton, H. Lee, K. McDonald, N. Murphy, N. O'Connor, and S. Marlow. Mobile Access to the Fischlár-News Archive. In *Proceedings of Mobile HCI 03 Workshop on Mobile and Ubiquitous Information Access*, pages 1-10, Udine, 2003.

[10] K.-M. Lee. Neural Network-Generated Image Retrieval and Refinement. In *Proceedings of the 1st International Workshop on Adaptive Information Retrieval (AMR 2003)*, LNCS series, Springer, 2003.

[11] X. S. Zhou and T. Huang. Relevance Feedback in Image Retrieval: A Comprehensive Review. *ACM Multimedia Systems Journal, Special Issue on CBIR*, 8(6):536-544, 2003.

[12] J. Urban, J. M. Jose and C. J. van Rijsbergen. An Adaptive Approach Towards Content-Based Image Retrieval. In *Proceedings of the Third International Workshop on Content-Based Multimedia Indexing*, Rennes, France, 2003.

[13] A. Nürnberger. User Adaptive Categorization of Document Collections. In *Proceedings of the 1st International Workshop on Adaptive Information Retrieval (AMR 2003)*, LNCS series, Springer, 2003.

[14] A. F. Smeaton. Challenges for Content-Based Navigation of Digital Video in the Fischlár Digital Library. In *Proceedings of CIVR-2002*, London, 2002.

[15] H. Lee and A. F Smeaton. Searching the Fischlár-News Archive on a Mobile Device. In *Proceedings of the ACM SIGIR 2002, Workshop on Mobile Personal Information Retrieval*, pages 32-41, Tampere, 2002.

[16] P. Browne, C. Czirjek, G. Gaughan, C. Gurrin, G. J. F. Jones, H. Lee, S. Marlow, K. McDonald, N. Murphy, N. E. O'Connor, N. O'Hare, A. F. Smeaton and J. Ye. Dublin City University Video Track Experiments for TREC 2003, IN *Proceedings of the TRECVID 2003 Workshop*, Gaithersburg, MD, USA, 2003.