Aggregated Feature Video Retrieval for MPEG-7 via Clustering

Jiamin Ye, B.Sc., M.Sc.

A dissertation presented in fulfilment of the
requirement for the degree of Doctor of Philosophy

Supervisor Prof Alan F Smeaton

School of Computing
Dublin City University
Glasnevin
Ireland

August 2004

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work

Signed _____

ID No 5016 2365

Date 14 Sep 2004

This thesis is dedicated to my parents, Koon-Kei Yip and Yee-Lan Wu,
who continue to support my taking on new challenges

ACKNOWLEDGEMENT

I especially would like to thank my husband, Paul Browne for his support during the course of my study and his patience and willingness to review the complete thesis multiple times, providing invaluable suggestions on its ideas and contents

I sincerely thank my family for their encouragement and understanding, who are always there to share my joy and misery with Also thanks to my uncle and auntie for their support and advice on my daily life

I kindly acknowledge my supervisor, Prof Alan Smeaton for his generousness of giving me the opportunity to pursue the degree, his guidance and suggestions during the course of my study and his invaluable advice on the ideas and contents of my thesis

I also would like to thank the people in the old postgraduate lab, Shuhong, Ling, Yongle, Wu-hai, Jer, Aidan, Thomas, Saba, Atid, and the people in the new CDVP lab, Paul-Browne, Kieran, Hyowon, Cathal, Georgina, Neil, Sandra, Paul-Ferguson, Pete and Dr Lao Special thanks to Cathal for his kindness in providing me a fast machine to run my experiments Many thanks to Paul-Browne, Kieran and Wu-hai for their support and the fun I had during the afternoon coffee breaks

I kindly acknowledge my previous supervisor, Dr Mark Sanderson for his guidance and encouragement of my taking up such a precious opportunity I was given

I gratefully acknowledge the support of the Informatics Directorate of Enterprise Ireland for my studentship funding from both the IP2000 and L'OEUVRE projects

Thanks to the generous funding of DELOS for my attendance at DELOS 2001 summer school on digital library technologies and the funding of CEPIS-EIRSG and of AICA for my attendance at ECIR 2003

Finally, I would like to thank School of Computing, DCU, for providing me a friendly and enthusiastic research environment

ABSTRACT

MPEG-7 is a generic standard used to encode information about multimedia content and often, different MPEG-7 Descriptor Schemas are instantiated for different representations of a shot such as text annotations and visual features Our work focuses on two main areas, the first is devising a method for combining text annotations and visual features into one single MPEG-7 description and the second is defining how best to carry out text and non-text queries for retrieval via a combined description

We align the video retrieval process to a text retrieval process based on the TF*IDF vector space model via clustering of low-level visual features Our assumption is that shots within the same cluster are not only similar visually but also semantically, to certain extent Our method maps the visual features of each shot onto a term weight vector via clustering This vector is then combined with the original text features of the shot (i e ASR transcripts) to produce the final searchable index

Our TRECVID2002 and TRECVID2003 experiments show that adding extra meaning to a shot based on the shots from the same cluster is useful when each video in the collection contains a high proportion of similar shots, for example in documentaries Adding meaning to a shot based on the shots that are around it might an effective method for video retrieval when each video in the collection has low proportion of similar shots such as TV news programmes

TABLE of CONTENTS

CHAPTER FIVE

CHAPTER SIX

CHAPTER SEVEN

LIST of FIGURES

CHAPTER SIX

CHAPTER SEVEN

LIST of TABLES

CHAPTER ONE

CHAPTER TWO

CHAPTER THREE

CHAPTER SIX

CHAPTER SEVEN

Chapter One

Digital Video Retrieval

People actively search, gather, share and consume information everyday When faced with an information need, most people may first turn to friends for help, if that is not available, a search in a library or on the Internet may be carried out Once the relevant information is found, a detailed examination of the content can follow (though this is not always the case) In some cases the initially selected information might be enough for the existing need If not, additional information may be gathered Having digested the collected information, some people need to re-organise the information for sharing with others in a class/lecture or on the Internet Rapid access to information of all kinds has become an integral part of our life for businessmen, administrators, researchers, lecturers and students alike Providing people with an independent ability to access and understand information has been the aim of present-day efforts to improve productivity

Information retrieval is the subject of searching for information in documents and mainly deals with the representation, storage and access to documents The documents of interest can be in any forms such as metadata, text, XML, images sounds or videos, each of which has its own bodies of structure, theory, algorithms and technologies

An information retrieval system whether it is manual or automatic can tell users the present or absent and locations of documents associated to their search requests In the context of text retrieval, the input information to the system is often the text keywords from documents (and abstracts) and the output corresponding to a request consists of a list of items The list, probably ranked based on the similarity between

the documents and the request, is intended to be shown to users for relevance judgements

One problem of finding relevant items is the possible different vocabulary used between the authors and the user One solution is to incorporate a thesaurus into the retrieval system to include potential related terms for a user's initial query Another popular solution is to employ relevance feedback to reformulate new queries by using the relevant items previously judged by the user The assumption of relevance feedback is that the new queries show a better degree of similarity with the previously identified relevant items than the original queries and the new queries will improve future searches [Salton & McGill, 1983]

The major advantages of information retrieval systems are that (1) they accept natural language requests and incomplete specification of the requests is allowed as opposed to complete query specification with restricted vocabulary and syntax such as those required in a database search, (2) information retrieval systems search for relevant documents as opposed to exact matching documents [Rijsbergen, 1979]

The availability of modern information retrieval systems has provided increased capabilities for the management of all kinds of information and greatly improved the access to many stored information collections Web search engines such as Google[1] and Yahoo![2] are amongst the most popular implementations of information retrieval research

Digital video retrieval is a subset of information retrieval and the main form of information that a video retrieval system handles is video sequences digitised in the uncompressed or compressed format Consider the problem of a user locating a particular segment in a two-hour long video Video is often broken down into shots for indexing and easy access purposes, where a shot is defined as a continuous video sequence recorded by a single camera without interruption Shot segmentation is the problem of detecting the boundaries between consecutive shots and the approach in general is to define a suitable threshold which represents significant differences

[1] Google, http //www google com/
[2] Yahoo!, http //www yahoo com/

between adjacent frames in terms of some primitive visual features such as colour, texture and shape [Browne et al, 2000]

A key frame, usually the middle frame in a shot, is chosen to represent the shot and serves two different purposes Firstly, the collection of such key frames presents a shortened representative version of a given video segment and users are not required to watch the whole video during browsing Secondly, a shot can be effectively indexed by extracting primitive visual features from its key frames thereby reducing the enormous efforts required in indexing features from every frame of the shot

A video consists of audio and visual information Content indexing can be accomplished based on both spoken text detected from the audio and visual features extracted from key frames Different access methods can be provided accordingly, relating to either the semantic or the perceptual aspect of videos A search for video content is often complemented by text and image-based queries In short, video retrieval is a research effort to develop methods of utilising both the existing text and image retrieval technologies

In section 1 1, we present some statistics figures about the amount of information produced in 2002, in particular video content, and describe the needs of digital video retrieval The principle components of a digital video retrieval system will be given in section 1 2, namely the representation of video content and structure, indexing methods, video similarity and query formulation and visualisation In section 1 3, we briefly introduce three main experimental video retrieval systems and they are the Informedia project from Carnegie Mellon University, CueVideo from the IBM research group and the Fischlar project from Dublin City University Key problems encountered in video retrieval and future research directions will be discussed in section 1 4 Finally, we will give the organisation of our thesis in section 1 5

1.1 The Needs and Challenges of Digital Video Retrieval

The amount of information produced and stored in different physical storage media (i e print, film, magnetic and optical) has increased dramatically over the last 10 years A study conducted by the School of Information Management and Systems at the University of California at Berkeley gave a rough estimation for how much original content (not copies) is created by the world every year, if stored digitally [Lyman & Varian, 2003] They estimated that there were approximately 2 to 3 exabytes[3] of new information created in 1999 and 3 to 5 exabytes in 2002, a growth rate of 30% per year between 1999 and 2002

Table 1-1 World-wide stock of video original content in 2002, if stored digitally in terabytes

Storage Medium	**Type of content**	**Annual production Upper bound**	**Annual production Lower bound**	**Accumulated stock**
Film (analogue storage)	Motion Pictures (e g Cinema & TV programmes)	25,254	19,187	740,700 (from 1895 to 2002)
Magnetic (digital storage)	Video tape & MiniDV	2,605,000	2,605,000	15,425,000
Optical (digital storage)	DVD	43 8	43 8	88
	Total	**2,630,298**	**2,624,131**	**16,165,788**

Of the 5-exabyte information produced in 2002, approximately 2 6-exabyte information came from analogue and digital video content and the total amount of video content made around the world (from 1895 to 2002) was about 16 exabytes (see Table 1-1) The amount of new original video content stored on film was estimated to be 25 3 petabytes and it consists of all film production for public release, including films made for television The measure is based on 1 terabyte for recording one hour of motion pictures in high-quality archival storage For magnetic medium (i e video tapes and MiniDV), estimates were considered restrictedly to films made by individuals for private consumption and the amount of new video content is 2,605 petabytes based on 2 gigabytes for one hour of video content in the

[3] 10^6 bytes = 1 Megabyte 10^9 bytes = 1 Gigabyte 10^{12} bytes = 1 Terabyte 10^{15} bytes = 1 Petabyte 10^{18} bytes = 1 Exabyte 5-exabyte information is equivalent to all words ever spoken by human

MPEG-2 compression format The amount of new video content on optical medium (i e DVD) was 43 8 terabytes

The task of finding the desired video content accurately and quickly becomes much more difficult as video content keeps growing A need for automatic content-based indexing and retrieval of digital video has been seen

Video retrieval systems differ from image and text retrieval systems in that video content is a blend of sound and pictures Words being spoken over a video sequence enable a high level of video abstraction Even though text is semantically rich, a search based on key words might not be enough partly in that some videos or segments are not accompanying any words relevant to text queries

A search for video content is therefore complemented by image-based queries, addressing perceptual features of users' needs The innovation of computerised image analysis techniques and pattern recognition algorithms allows for the representation of perceptual visual features of video content such as colour, texture, shape, motion, objects and events These analysis techniques study the distributions and spatial localisation of pixels in a frame and the spatio-temporal relationships in a sequence of frames

However, image-based searches can often produce unexpected results due to the existence of a gap between the primitive visual features and the higher level cognition A photo of a sunset by the sea, for example, is characterised with large amount of redness and a search with this photo would also retrieve shots of a sunrise or even a fire scene

In order to bridge the semantic gap, studies have been carried out to identify semantic components/objects in videos and proposed video abstractions (or concepts) that offer enhanced textual descriptions about video (e g "car", "face", "landscape/cityscape") beyond visual primitives [Smeaton et al, 2003] The abstraction/annotation process is done semi-automatically via a training procedure Even though there are some visual features that can be modelled as semantic

concepts (i e things), there are a number of concepts for which the creation of a model is impossible (i e feelings and ideas)

In short, the growing amount of video content has attracted research efforts in providing tools for effective video retrieval beyond simple key words - based on primitive visual features Additional burden is required for handling video information, which extends its text form by including motion pictures and audio, making the information more rich and expressive, but at the same time more complex to process by computers Contemporary video retrieval systems are unable to deal effectively with the increasing growth of video information In spite of many efforts on the research, current operational capabilities remain at a relatively elementary stage partly due to the complex video matching techniques

1.2 Principle Components of a Digital Video Retrieval System

Access to digital video involves many issues including the gathering, processing, searching of data and result presentation to end-users The primary concern of digital video has been for storage and transmission Video compression standards have focused on ways to maximally reduce the amount of data that can be stored or transmitted without too much loss of picture and sound quality, but not on how the compressed data can be used for video structure and content analysis effectively

The standards assume that the compressed data will be decompressed before subsequent analysis and as a result the cost for decompression increases as the amount of video required for analysis grows A great deal of studies have been carried out to directly use the information available in compressed data and have shown that results obtained in such a way are comparable to those obtained via decompression [Calic et al, 2002] [Yeung et al, 1996]

Another concern of digital video is the need for representations of video content and structure to assist interoperability In October 1996, the MPEG group started a new work item called MPEG-7 to provide a solution to the representation problem

MPEG-7 is a metadata-based standard, addressing information about the multimedia content

For content-based analysis, MPEG-7 is flexible enough to allow representations at various levels from objects, to frames, to video structures For content-based indexing, it provides generic descriptions for various features and the features are designed in such a way that they are easy to extract and have meanings in the context of different applications For visualisation and browsing, it allows for an efficient way to display the summary of video sequences rather than a linear scan of each key frame

In the remainder of this section, we focus on the video representations and describe four basic components are required to develop a video retrieval system in turn (1) representation of video structure and content, (2) indexing methods, (3) video similarity and (4) query formulation and visualisation

1.2.1 Representations of Video Content and Structure

A video conveys information through multiple channels These include the way in which shots are connected together logically by using story plots and physically using editing effects such as cuts, fades, dissolves The segmentation of video into shots has been widely addressed and changes in colour, texture, shape and motion can assist the detection of shot transition and shot grouping

But the identification of story structure is more difficult and should be made to use genre-specific knowledge Each video genre (commercials, news, movies, documentaries, and sports) has its own particular characteristics and these are reflected in the way in which video units are extracted and organised in knowledge structures, indexed and accessed by users For example, a typical structure of a TV news programme consists of news headlines and stories News headlines are presented in the beginning of the programme to introduce the major stories to come Each story is self-contained and has three basic parts the opening, the narrative and

the commentary Topic changes in spoken words [Stokes et al, 2004] or news anchor person detection [Czirjek et al, 2003] can help identify news stories

We look at the structural representation of video from high to low level

- *Scene* – It is defined as a group of consecutive semantically related shots Television news stories are a form of scene If a video can be segmented into scenes, a user can browse through the video on a scene-based rather than a shot-based representation The main advantage is that a significant amount of information required to present to users is reduced, making it easy for users to identify information of interest The detection of scene boundaries is genre-dependant and current technologies are limited to segmenting news stories because of its well-structured and predictable story layout in news programmes

- *Shot* – It is defined as a continuous sequence of frames captured by one camera without interruption The shot boundaries are relatively easy to detect computationally due to the mechanical way of film editing (i e changes in colour, texture and shape between subsequent shots) In most video retrieval experiments the shot is seen as the minimum retrievable unit for users partly because it contains the minimal semantic information perceivable by users

- *Key frame* – It is a frame selected to represent a shot If there is some motion in a shot, we can simply obtain the middle frame in the shot sequence as the key frame If there is much motion in the shot, multiple key frames are preferable When a shot is represented by one or more images and the problem of video retrieval is reduced to the problem of image retrieval, for which video sequences are treated as collections of still images where traditional image retrieval techniques can be directly applied

Figure 1-1 below shows the hierarchical structure of a sample video sequence. The sequence is divided into two scenes: the development of a city centre and a tram approaching. To describe the developed city centre, two shots are used: one to show the skyscraper and the other to give a view of the centre from the harbour. These two shots can be detected by the changes of colour and texture between frames. Finally, middle frame is selected to represent each shot.

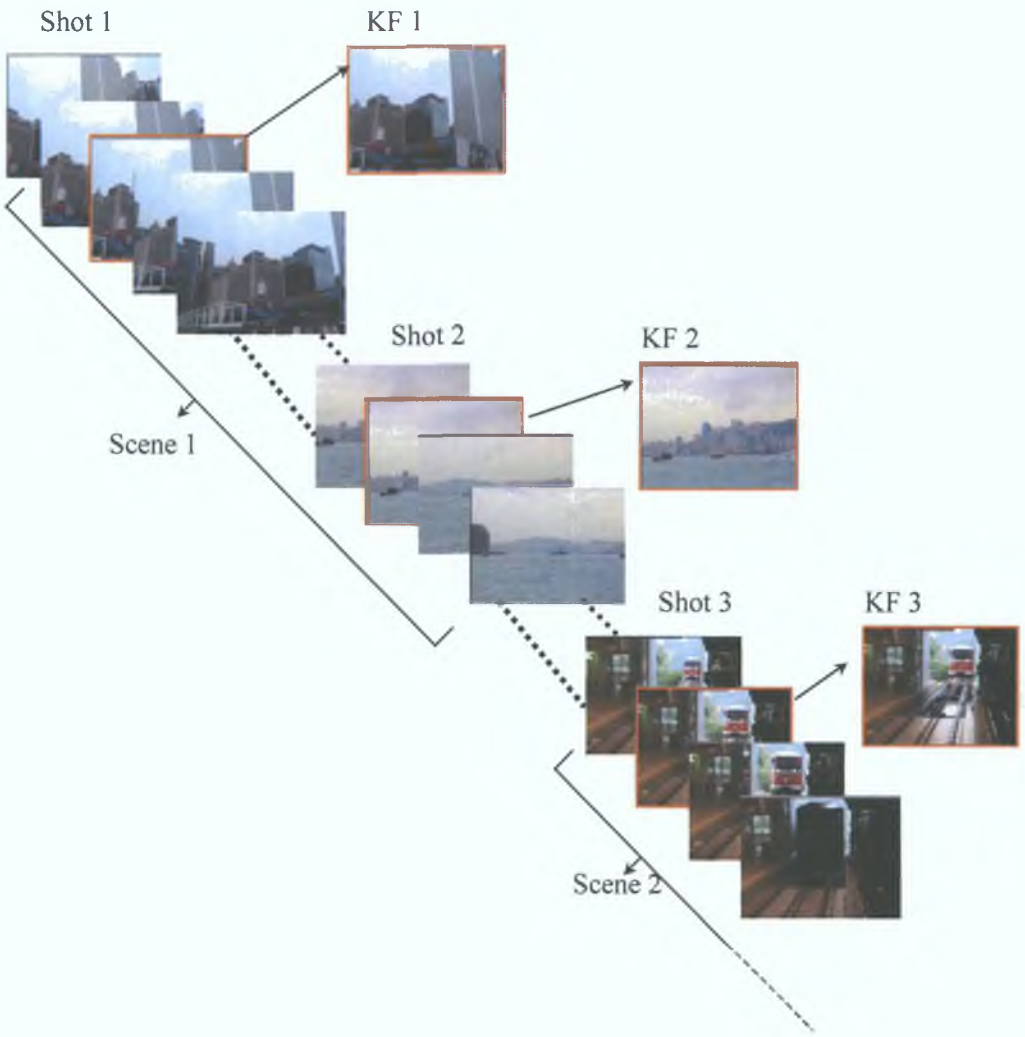


Figure 1-1: The hierarchical structure of a sample video sequence

The content of a video can be represented at many levels of abstraction (i e scene and shot) and we focus our study at the shot level Three types of content can be associated with shots

- *Primitive visual features* – They are the low-level visual features such as colour, texture, shape and motion Colour, texture, shape can be computationally calculated for a key frame or regions of a key frame Motion is used to characterise the camera or object movement over a sequence of frames The problem of using these features for video retrieval is that they lack semantic information Conventional techniques can help find limited things, for example that have similar colour and shape

- *Concept-based features* – They are the semantic labelling of video shots Each shot is associated with a mid-level concept label (i e "indoor/outdoor", "aircraft") that provides a brief description of the content of the shot Automatic concept detection takes advantage of the similar representations of shots that contribute to the same concept label and learns a generic representation of the concept via a training process [Smeaton et al, 2003] Given a new shot outside the training set, if the shot's representation and a concept representation are a good match, the shot is assigned to the concept The selection of concepts determines things that can be represented and retrieved within a video

- *Text features* – They are the text spoken over a shot and can be obtained via four different sources (1) subtitle/teletext signal from TV cable, (2) Automatic Speech Recognition (ASR) outputs, (3) subtitle information from a DVD, (4) transcripts of some TV programmes made available on the web In general the transcripts or textual descriptions are considered to be the best features used for video retrieval in that they contain higher level of semantic information than that of the mid-level features These transcripts normally come from ASR but not always

In Chapter 2, we will give more details of the MPEG-7 standard for representing video content and structure and the related retrieval methods for video

1 2 2 Indexing Methods

Indexing is a task to assign appropriate identifiers to shots capable of representing the content of shots and providing fast search functionality while assuring all items satisfying a user's query are returned with a high ranking In the context of similarity matching for text features, indexing is generalised in two steps (1) to assign terms capable of representing the content to each item, (2) to estimate the weight for each term whose value reflects the importance of the content identification purpose The terms used for video content representation normally come from the ASR text associated with the shots In the context of similarity matching for visual features, each shot is assigned to a numerical vector that compactly represents the features of the shot

For a small search collection, a sequential scan of all items followed by straightforward similarity computations is adequate But as the collection grows, this technique can be too slow and the conventional solution uses a multidimensional spatial access method such as R-trees to index each vector representation as a point in a *n*-dimensional Euclidean space [Guttman, 1984] The idea of R-trees is to represent a recursive subdivision of the n-dimensional space using a height-balanced tree Each non-leaf node in the tree represents a splitting subspace spatially containing at least one data point and a minimum bounding box is used for the subspace representation Non-leaf nodes have one or more descendants and leaf-nodes contain location pointers to data items in the collection Given a query point in the n-dimensional space, a search is carried out in a top-down fashion to locate the possible paths based on the criterion that the query point is contained in the box But the search may go down many paths because the bounding boxes may overlap

Spatial access methods work well when the number of vector dimensions is low but the computational cost increases exponentially with the dimensionality Two popular approaches to the problem have been used to speed up the search filtering and dimensionality reduction

- *Filtering* – A filter is seen as a computationally fast representation defined in a lower dimensional space for all items in a collection A search proceeds in two stages using two different representations (1) to compute the similarity between items and a given query in the low dimensional space, (2) items that pass through a threshold are further examined by using the original full representation In the QBIC system, retrieval based on colour features is carried out by first filtering the set of images based on their average RGB colour in a 3-dimensional space, then a more precise matching using a 256-dimensional colour histogram [Faloutsos et al, 1994]

- *Dimensionality reduction* – Dimensionality reduction extracts useful information from a feature in the original high dimensional space to form a new representation in a low dimensional space, as low as two or three dimensions A search can be performed on such a transformed space Discrete Fourier Transformation is one of the popular dimensionality reduction methods used in image retrieval The idea is that any signal can be represented as a linear combination of sine and cosine waves Each wave is associated with particular frequency and has a corresponding Fourier coefficient (i e weight) in the linear combination [Press et al, 1989] The first few coefficients after the transform normally carry most of the information about the original signal and they can be used for approximating the signal

  For text searching, Deerwester et al describe the Latent Semantic Indexing method using the Singular Value Decomposition technique [Deerwester et al, 1990] The idea is that the occurrence of some patterns of words can give evidence to the likely occurrence of others based on the observation that different people use different words to describe the same concept An index is created by first computing the singular values (i e orthogonal factors) of a large term by document matrix, then choosing the best singular values to represent the important "concepts" in the original matrix The derived concepts rather than the original terms are used for retrieval

1.2.3 Video Similarity

The judgement of similarity is related to human subjective and required to closely conform to human similarity perception A well-known theory assumes that such perception is based on the measurement of an appropriate distance in a metric space such as Euclidean distance Comparing two video sequences (or shots) requires cross-matching of individual segments from the two sequences and a simple distance measure between points in a multidimensional space may not be sufficient Liu et al defined four criteria to measure the similarity between videos (1) visual, (2) temporal order, (3) temporal duration, and (4) granularity similarity [Liu et al, 1999]

- *Visual* – Two videos are similar if they both present similar visual features such as colour, texture, shape and motion The features are usually derived from a set or a sequence of individual frames or shots and each feature may have its own distance measure function

- *Temporal order* – It is the ordering constraints on the appearance of frames, shots, or scenes Some shots have the same visual features but could be arranged to appear in different temporal order within different videos This similarity is generally examined in the final stage of measurement

- *Temporal duration* – It is the discrepancy of the same video content on the speed of temporal development Some shots may present different length in different videos For instance, a shot may have a short temporal duration in an edition where the temporal development is required to be fast while a long temporal duration indicates a slow temporal development

- *Granularity* – It focuses on the hierarchical structure correspondence between videos Ideally two videos are considered similar when most shots and scenes in one video can find similar ordered correspondences in the other

The last two criteria about the temporal duration and granularity are not necessarily restrictive due to the diversities of temporal duration and hierarchical structure correspondences used in the video editing process [Chen & Chua, 2001]

For visual similarity, current image matching method can be directly applied each shot in a video can be represented by one or more key frames as stated in section 1 2 1 and shots may be compared by any visual features derived from the key frames

For temporal order similarity, approaches generally consider the number and order of good frame matches between two shots, while ignoring those not matched [Lienhart et al, 1998] [Chen & Chua, 2001] and a predefined threshold is given to decide a good frame match (or one-to-one correspondence)

[Lienhart et al] proposed a re-sequencing measure that determines the number of minimum relative re-orderings required to transfer the sequence of shot *A* into a new sequence which has the same ordering of frames as that of shot *B* It is defined as a ratio between the number of re-orderings and the total number of good frame matches between two shots The ratio has a range between 0 and 1 A low re-sequencing measure value indicates that the temporal characteristic of shot *B* is preserved well in the shot *A*, namely less re-orderings required transforming shot *A* to *B*

Chen et al used a sliding window approach to locate similar video sub-sequences to a given video clip [Chen & Chua, 2001] The approach slides a window over shots of a video by one-shot increment, calculates a ratio between the number of good frame matches within the window at each shot position and the window size, and finally plots the set of ratio values in a curve graph The curve graph shows the degree of similarity between a query video clip and a video sequence in the collection A local maxima indicates the beginning of possible similar video sub-sequence Given a threshold, some local maxima can be removed, where the number of matching correspondences is not significant enough to include the original video m the final result

The above four criteria for video similarity are concerned with the physical similarity among video sequences and subject to the different ways of film editing However, another aspect of video similarity is overlooked logical and semantic similarity Logical similarity refers to the organisation of stories in such way that their ordering appearance gives audiences the capability of reasoning and audiences try to interpret the stories and make sense of the events in the story These stories are not necessarily arranged in a sequential order but more often appear to be interweaving with others It is on the logical level that audiences must focus their attention, remembering that every event in the stories is connected to the development of the stories Current technologies are not able to detect multiple stories that occur simultaneously in a video

Semantic similarity basically studies the meanings of shots and scenes, namely the language used to achieve a desired effect on audiences during their understanding of the stories, especially through the use of words The classification of changes in the signification of words from the spoken texts is a popular way of comparing the semantic similarity between two video sequences But the use of text features may not work well in situations where words have dual or novel meanings

1.2.4 Query Formulation and Visualisation

The indexing and retrieval processes provide videos or shots satisfying users' information need But users are still required to issue their requests in a format that can be understood by a machine and to tediously view the returned results in a linear fashion Query formulation and visualisation mechanisms for digital video are subsequently developed to aid users in finding the relevant information

Bolle et al summarised the video formulation process as an iterated sequence of stages navigating, searching, browsing and viewing [Bolle et al, 1998] Each stage is seen as an information filtering process to reduce the size of candidates by appropriately using different types of video features and visualisation can then be designed accordingly

- *Navigating* – Metadata is the main form of video organisation in this stage Most video is associated with some metadata such as bibliographic, subject, genre information Users can decide which category of video to start the search with a specific genre (sitcom, news, documentaries, reality TV), a specific interval of time, or a specific topic (earthquake, universe, animal, plant) Users sometimes might need to select more than one category to search if certain video is assigned to multiple categories A result list is displayed and each item in the list contains the metadata of the candidate video

- *Searching* – Queries in the searching stage are issued based on text, audio, visual features of video or any combinations Searching is a challenging task since many aspects of video cannot be described simply in text or by keyframes Text search is a mature and straightforward technique also an important tool in the video retrieval problem Audio and visual features go beyond text and the difficulty is to define the appropriate features that extractable and distinguishable to trim down the candidates Query-by-image techniques have been widely used in video retrieval and similarity between candidates and a given query image is based on the image features such as colour, texture and shape Query-by-video-clip addresses much richer features of video such as camera motion, object movement and temporal duration Shots or segments rather than an entire video are returned in the relevant result list

- *Browsing* – This requires a good summary representation of candidate video Users can understand the video content easily by looking at the summary and can go through many video candidates in a short time

  At the shot level, visual summary might be thumbnail representation of keyframes of shots displayed linearly At the story/scene level, visual summary might be the hierarchical video display which breaks down a video stream into a number of stories, which are in turn broken down into a number of shots and each story/shot representation is thumbnail-based [Zhang et al, 1995]

At the video level, keyframes are selected as few as possible to adequately describe the video content such as the "filmstrips abstraction" tool developed by the Informedia project [Christel et al, 1997] Users can click on any point in the summary to see the video and visual summary provides a way for random access to candidate videos

Video summarisation depends on the video genre For sitcoms, news and films, story structures are the major way of organising the summary For a sporting event, the summary focuses on detecting actions such as scoring, defence and offence [Sadlier et al, 2003]

- *Viewing* – users can view any parts of the selected candidates by simply clicking on the thumbnail images in the video summary representation The basic functions that support the video playback are play, pause, fast-forward and fast-rewind

1.3 Review of Three Digital Video Retrieval Systems

We review three contemporary digital video retrieval systems in turn by focusing on their functionality and characteristics (1) the Informedia system, (2) CueVideo and (3) the Fischlar system Although experimental video retrieval systems existed at the beginning of the 1990's, the Informedia system would be considered as the start of automatic video retrieval systems Most available video retrieval systems are from academia It would be difficult to name or compare them all but some well-known examples include CueVideo and Fischlar that all provide video analysis, indexing and retrieval functions

1 3 1 The Informedia System

The Informedia system at Carnegie Mellon University, initiated in 1994, is a comprehensive and complete digital video retrieval system which provides search and retrieval of current and past TV, news and documentary programmes The system currently contains about 2,000 hours, 1 5 terabyte of video that has been automatically processed by integrated speech, image and natural language technologies [Wactlar, 2000] The group helps open up a new line of thought and technical development for automated video and audio indexing, navigation, visualisation, search and retrieval

The system provides fast and accurate text access to video content based on a combination of the words spoken in the soundtrack, keywords overlaid on the screen images and available text obtained during close-captioning Also supported is the matching of similar faces and images The latest work attempts to create summarisation and visualisation of video content from text, audio, image and video into one single abstraction Map-based display is incorporated into the system to provide the historic and geographic perspectives of videos

1 3 2 IBM Research Group

The CueVideo project[4], established in 1997 in the Visual Media Management Group at IBM's Almaden, is a past project to study the indexing and retrieval of digital video collection Technologies being incorporated in the system to support a full range of video applications include audio/visual analysis, speech recognition, information retrieval and artificial intelligence The system supports different visualisation mechanisms such as moving storyboards and smart fast video and audio browsing

[4] The CueVideo project http //www almaden ibm com/projects/cuevideo shtml, last visit on 20 April 2004

Further research on digital video analysis is carried out by the Pervasive Media Management Group at IBM's T J Watson from 2001 till present The MARVEL MPEG-7 video search engine[5] is developed to demonstrate the complex indexing and retrieval technologies Much research efforts are put on the model-based analysis and search fusion methods [Amir et al, 2003] The model-based analysis aims at automatically classifying video content and assigning concept labels (i e indoor/outdoor) to each shot in the archive using statistical models The search fusion method assists the combination of results generated separately from content-based retrieval (i e based on primitive visual features), model-based retrieval, text- and speech-based retrieval engines, allowing users to construct different types of queries

1.3.3 The Físchlár System

The Fischlar system was started in 1997 by the Centre for Digital Video Processing group at Dublin City University by providing advanced access to broadcast TV content [Smeaton et al, 2001] The group developed technologies for automated recording, browsing, searching, alerting and summarising video content The system supports full text retrieval at various granularities, from shots to segments/stories Each story is automatically linked to related stories or video from the archive based on the similar spoken text patterns Fischlar has been applying different summarisation and visualisation technologies for four different video collections to assist users' navigating, searching, browsing and viewing Fischlar-TV, Fischlar-News, Fischlar-Nursing, and Fischlar-TREC2002

Figure 1-2 below is a screen shot of the Fischlar-News system On the left-hand side of the screen, a calendar facility is provided for users to fast access to the daily RTE1 9 o'clock news archive dating back to April 2003 A text search function is provided in the system and a query of "presidential election" is given in the search box on the top-left corner of the screen A list of 194 relevant stories with the corresponding anchorperson icon and text summary is returned on the right-hand

[5] The MARVEL MPEG-7 video search engine http //www research ibm com/pmm/projects html, last visit on 20 April 2004

side of the screen and they are sorted based on chronological order. Clicking on an anchorperson icon on the result list, a user is led to the corresponding detailed story including the keyframes and subtitle information of the complete story. A playback option for each news story is also provided for users' viewing.

Current research in the group focuses on the studies of advanced video analysis (object, event detection), MPEG-4 encoding, MPEG-7 representation and subsequent video retrieval technologies.

Figure 1-2: A screen shot of Físchlár-News

1.4 Research Directions in Digital Video Retrieval

Current operational capabilities of most video retrieval systems remain at a relatively elementary stage partly due to the complex video matching techniques. Text search is a mature and straightforward technique and also widely used in video retrieval along with natural language understanding techniques. But there is often a case where text does not give us enough information about what is seen in the shots and we therefore need the use of low level, automatically obtainable visual features. The

chosen visual features determine the characteristics and retrieval capabilities of the video retrieval systems

As the size of video collections increases, the need to model semantics via concept labels as patterns of visual abstractions across the collection become evident [Smeaton et al, 2003] The meaning of the label is concerned with logic relationships between objects/events and the domain that is represented The relationships are designed by certain stable categories of ontology such as animal, plants, scenery and human Each shot is associated with one or more labels A computationally fast filter can be applied to all shots and only shots that have the corresponding label are operated on in the second stage of searching However the conceptual representation of visual content is studied with a limited number of labels, and research in this direction is still in its infancy

The MPEG-7 standard for representations of video content and structure has yet to be completed and new techniques are continuing to be tested The development of the standard requires that all proposed descriptors (i e representations) must be meaningful in the context of different applications The descriptors are chosen in extensive tests and their various characteristics are assessed before being adopted in the final standard [Sikora, 2001] A good descriptor captures the structural, visual and textual properties of a video in a concise way and renders itself to fast searching and browsing

With the introduction of MPEG-7, a description of a video file can be generated and used by search engines New generations of search engines will be required to interpret the image or video clips that we are interested in and compare it to all the files on the Web, much as we do today for Web pages The area for video retrieval using MPEG-7 descriptions is still young and operational search engines have yet to be developed The task in is thesis is set in this perspective

An MPEG-7 description is equivalent to a XML document if only text features are considered But access to MPEG-7 is different from the XML-based approaches because MPEG-7 considers image-based features that rely on the characterisation and pattern matching of visual primitives The interest of this work focuses on a

method to integrate visual information into an available XML retrieval approach and our approach will be given in Chapter 4

Up to 2001 the evaluation of video retrieval systems was carried out mostly by academics using a few small, well-known video collection corpora, and their own manually annotated content [Browne et al, 2000] or even smaller test sets such as the MPEG-7 video content set [MPEG7-W2466] In 2001 the annual Text Retrieval Conference (TREC) sponsored by the Defence Advanced Research Project Agency (DARPA) and the National Institute of Standards and Technology (NIST) created an annual workshop (TRECVID) dedicated to the research into digital video retrieval The TRECVID workshop provides a common base for retrieval system experimentation and comparison by supplying standardised evaluation, test collections and query topics A great deal of experience has been gained on TRECVID since the first running of the evaluation track in 2001, as the details and focus of the video retrieval experiments have evolved We will test the performance of our approach by using the TRECVID2002 and TRECVID2003 collections, and the results will be given in Chapter 6 and 7, respectively

1.5 Organisation of the Thesis

The thesis contains eight chapters which can be divided into three topic areas (1) literatures on structured document retrieval technologies for MPEG-7 and XML, (2) our proposed approach on aggregated feature retrieval for MPEG-7, and (3) experimental setting and results on the approach

In Chapter 2, we present an in-depth review of the representations of digital video content and structure defined in the MPEG-7 standard Also given is the literature survey of the contemporary retrieval methods dedicated to MPEG-7 We conclude the chapter by comparing the pros and cons of the methods and indicate the need to explore the combined use of both text-based and content-based techniques for MPEG-7 searching

In Chapter 3, we extend our review of current XML document retrieval approaches in order to address their strengths and potential in terms of video retrieval We classify the techniques for XML document retrieval into three categories (1) Information Retrieval based (i e IR-based), (2) Path Expression and (3) Tree Matching We suggest that the aggregation-based approach in the IR-based category provide a better foundation for MPEG-7 video retrieval since they allow for content matching and approximate structure matching The assumption of the approach is that a document's structural information is additional evidence to the original content of XML elements

In Chapter 4, we propose an approach to model both textual and visual representations of MPEG-7 descriptions in Chapter 4 We adopt the aggregation-based approach based on a similar assumption that primitive visual features are auxiliary information to the original semantics of video shots Our solution is to align a video retrieval process to a text retrieval process based on the TF*IDF vector space model via clustering of primitive visual features We map the visual features of each shot onto a term weight vector by their proximity to the corresponding cluster centroid and the terms are collected from transcripts of all member shots in the cluster This vector is then combined with the original text descriptions of the shot to produce the final searchable index

In Chapter 5, we present the literature of evaluations for information retrieval systems, particularly some popular performance measurements used in TREC experiments Our proposed approach was tested based on both the TRECVID2002 and TRECVID2003 collections and the details of experimental settings and results will be given in Chapter 6 and 7, respectively Chapter 8 concludes this thesis by summarising the evaluations of our experiments Some insights to the problems of the proposed approach are also studied

Chapter Two

Standards for Representing Digital Video Content

Text annotation, the manual annotation of some digital information by text "labels", is a popular method of describing image and video content In the earliest days of content based image retrieval research, image and video retrieval could be carried out by a combination of text based annotation and a text based database management system With more and more images and videos now becoming available, manual annotation for large collections becomes labour intensive As a result of this, automated content-based image and video retrieval techniques emerged where images could be described by their visual features rather than just text These visual features include colour, texture, shape of objects, etc Video clips could further be described by their hierarchical structure which can then be combined with their audio/visual features

In response to the need for the use of standards for multimedia content representation, the MPEG group put forward the MPEG-7 standard supporting metadata, textual annotation and content-based descriptions The goal of the MPEG-7 standard is to achieve the maximum interoperability, that is, to support the interchange of multimedia descriptions in unambiguous formats such that all consuming systems conformant to the MPEG-7 standard can understand the structure [MPEG7- N4039]

A detailed introduction to the MPEG-7 standard is given in section 2 2, including the components and specifications of the standard Sections 2 3 and 2 4 will detail the Multimedia Description Schemes and Visual Description respectively but first, in section 2 1, we give a review of existing MPEG standards

2.1 Existing MPEG Standards

The Motion Picture Experts Group (MPEG) was formed in the mid 1980's by the main multimedia industry leaders including Sony, Philips, Hitachi and scientists involved in digital content-based research There was a need at the time for a standardised method for display of digital video on computers At that time the CD-ROM technology was just starting to emerge and some agreed video format was needed that could be decoded on the popular and available computer hardware of that time

The MPEG-1 format, established in 1988, specifies a coded representation for compressing video and audio sequences to a bitrate around 1 5M bit/s It also addresses the functions of decompression and synchronisation of the audio and video data streams that are conformed to MPEG-1 [Chiariglione, 1996] Video CD a predecessor of DVD Video is a variant of MPEG-1

MPEG-2 was designed for high quality video and handles larger picture resolution It targets a higher bitrate environment from 3 to 10M bit/s [Chiariglione, 2000] Digital Satellite, Digital Video recorders and DVD Video disks are all products that use variants of MPEG-2

MPEG-4 is aimed at digital video for broadband Internet (ADSL, Cable Modems) and the Intranet The standard utilises advances in visual compression to reduce bitrate and filesize requirements still further while offering improved quality [Koenen, 2002] It has been designed with object-based decoding in mind, individual objects like a chair or a person could (theoretically) be encoded differently from their backgrounds Currently MPEG-4 encoders or indeed any image analysis systems, cannot do accurate object detection on natural video and as a result current MPEG-4 encoders treat each frame of video as an entire single object, but the technology of object detection and segmentation is being developed

2.2 Introduction to the MPEG-7 standard

MPEG-7, known as the Multimedia Content Description Interface, standardises the syntax and semantics of multimedia descriptions [Day & Martinez, 2002] The design of the descriptions takes into account various application requirements including content management, searching, navigation and user interaction and covers a wide granularity, from an image region, an image, a video segment, or a video, right up to a whole video collection MPEG-7 however does not address the implementation of any applications but simply is a standard for describing content

Unlike previous MPEG efforts of MPEG-1, -2 and -4, MPEG-7 is not a multimedia compression standard MPEG-1, -2 and -4 make content available while MPEG-7 is metadata based, addressing information about the content and should facilitate retrieval of required content The MPEG-7 standard contains four main components

- *Descriptors* (Ds) - the basic descriptions that define the syntax and semantics for visual and audio features These are reusable tags or structures
- *Description Schemes* (DSs) - the semantics and structure relationships between descriptions They can be any Ds, DSs or both
- *Description Definition Language* (DDL) – the language for creating new DSs or Ds and for modifying/extending any existing Ds or DSs
- *Systems tools* – the specification for the synchronisation and transmission mechanisms

Figure 2-1 illustrates the relationship amongst the four MPEG-7 components DDL is basically a standardised XML schema with charateristics dedicated to MPEG-7 providing ways of defining and creating new datatypes (such as matrix, time and location) and new structured elements Both the MPEG-7 schema and the extension of existing Ds and DSs are defined in DDL The MPEG-7 schema models the hierarchical structures of the DSs and Ds An MPEG-7 document is an instance of the hierarchical model and is validated by the MPEG-7 schema Finally, MPEG-7

documents can be compressed into a binary format for efficient transmission and streaming if necessary

MPEG-7 bases representations on XML but XML has not been developed to use in any real-time environment like in the multimedia or mobile industries The overhead introduced by having to process a small amount of mark-up tags is not critical for parsers in most cases such as those for XML and HTML on the Web However, MPEG-7 is designed as a language for encoding audio-visual metadata and has high level of structural redundancy BiM (Binary Format for MPEG-7) is thus defined to facilitate the transmission and processing of textual representations [Martinez, 2003] BiM is schema oriented because the schema is known both by the encoder and the decoder The main advantage of this is to remove the schema and structural redundancy from the MPEG-7 textual representation in order to achieve high level of compression

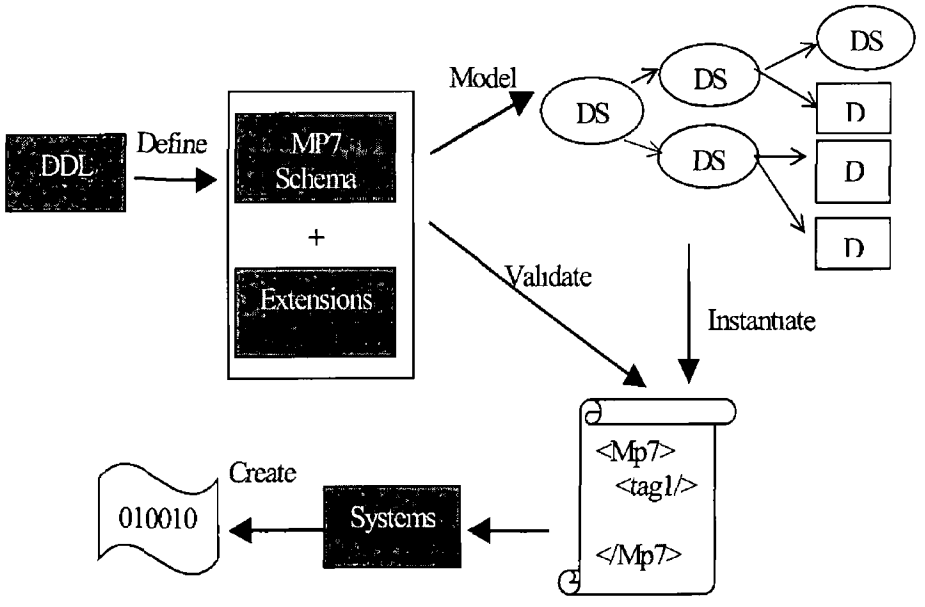


Figure 2-1 Relationship amongst the four MPEG-7 components [Martinez, 2003]

Of the four MPEG-7 components, two are of concern in our work Ds and DSs Ds consist of mainly the visual and audio part for annotating the outcomes from visual and audio detectors Section 2 3 describes the visual part in detail, in particular the colour and texture Ds DSs have a wider coverage of multimedia

aspects including structural and conceptual representations, creation and production, usage rights, user preferences and history, etc Section 2 4 gives insights into the structural representations of videos For more detailed introduction into multimedia DSs, the reader is referred to [Salembier & Smith, 2001]

2.3 The Representations of Visual Features

The MPEG-7 visual descriptors include five main types of visual features (1) colour, (2) texture, (3) shape, (4) motion, (5) face recognition For more details, the reader is referred to [Manjunath et al, 2001] Our work concentrates on the general visual features which can be easily extracted and used in most applications, namely colour and texture MPEG-7 defines five colour descriptors as listed below

- *Colour space* – Monochrome, RGB, HSV, YCrCb and HMMD are all supported in MPEG-7 The monochrome space corresponds to the Y component in the YCrCb space The transformation among the different colour spaces is possible and specified in [Cieplinski et al, 2001] RGB to YCrCb is a linear transformation RGB to HSV and HMMD are both non-linear and reversible conversions

- *Dominant colour* – It corresponds to a set of representative colours for a whole image or for arbitrary regions The representative colours can be from any type of colour space and are formed via a colour space quantisation process The quantisation process groups colours in a given region into a small and arbitrary number of representative colours Each representative colour records its percentages and colour variance within the region The maximum allowed number of representative colours is 8 and the minimum is 1 for each colour component in the colour space

- *Scalable colour* – It is defined as a colour distribution/histogram for a whole image or an arbitrary region Colour space is constrained to the HSV space which is uniformly quantised to 16, 32, 64,128 or 256 bins Matching between scalable colour can be measured using an L1 norm by the distance between the histogram bin values Scalable colour is useful

when only colour presence and their amounts are taken into consideration, regardless of their spatial arrangement

- *Colour structure* – It corresponds to both the colour distribution and structure of colours for an image or an arbitrary region It is designed to capture the structure difference of a given colour between two images or regions in cases where both images present identical colour distributions

- *Colour layout* – It is defined as the spatial distribution of colour of a whole image or of an arbitrary region The colour space is restricted to YCrCb The feature is extracted by dividing the region into 64 (8 × 8) blocks, selecting the dominant colour of each block and performing a transformation of the 8 × 8 matrix on each colour component independently Coefficients are selected from the transformed 8 × 8 matrix based on a zigzag scanning order These coefficients are then further quantised to form the colour layout of the region

Texture is interpreted as "the visual or tactile surface characteristics and appearance of something" in dictionaries A pattern on a fabric and an oil painting are good examples where image texture is pronounced whereas a clear blue sky or a pale coloured smooth wall are examples of a smooth texture Technically speaking, the appearance of a texture is given by the spatial distribution of some basic primitives in an image (i e basic texture elements or patterns) Two main texture descriptors described below are extracted from the luminance component

- *Homogeneous texture* – It is a perceptual description of texture with regards to *regularity*, *directionality* and *coarseness* It is suitable for discriminating images that have homogenous texture properties *Regularity* identifies the periodicity of the appearance of the basic primitives in an image and it could be assigned to any of the three values "regular", "slightly regular" and "irregular" *Directionality* gives the dominant directions of the texture and there are 6 directions that could be chosen from in steps of 30 degrees starting from 0, i e 0, 30, 60, 90, 120 and 150 If no directions can be found in an image, "no directionality" will be assigned *Coarseness* represents the frequency distribution of the

dominant directions and 4 values are used to summarise the coarseness "fine", "medium", "coarse" and "very coarse"

- *Edge component histogram* – It is the spatial distribution of five types of edges "vertical", "horizontal", "45 degrees", "135 degrees" and "non-directional edge" This is suitable for describing non-homogeneous texture images An image is divided into 4*4 = 16 regions and the frequency of each edge is calculated for each region The final product of this feature has a total of 16*5 = 80 histogram bins

The shape information from image objects provides a useful feature for similarity matching Unfortunately current image analysis systems cannot do accurate object detection on natural video and therefore pre-selected and/or pre-segmented objects based on pre-defined templates are used in the shape extraction process to create the final descriptors

Shape descriptors are required to be invariant to scaling, rotation and translation Shape information can be either 2-D or 3-D Real world objects are 3-D in nature and 2-D shapes are usually the projections of the real world objects onto an image plane in image/video retrieval For more detail of shape extraction techniques, the reader is referred to [Bober, 2001]

- *3-D shape descriptor* – This is designed based on a shape spectrum concept The shape spectrum is defined as the histogram of shape indices computed over the entire 3-D surface A shape index is defined as a function of two principal curvatures of a point on the 3-D surface and is used to capture information about the local convexity of the surface

- *2-D shape descriptor* – This can be divided into two categories based on the notion of similarity contour-based and region-based A contour-based descriptor expresses only the contour of the object's shape A region-based descriptor studies the spatial distributions of pixels of the entire shape region and can describe any shapes such as a simple shape with a single connected region or a complex shape that has "holes" (i e a small area containing pixels that have very different values from those surrounding the area such

as black and white) in the object or several disconnected regions The region-based descriptor captures the internal details of the shape in addition to the contour

- *2-D/3-D shape descriptor* – This supports the combination of the 2-D descriptors to represent the features of a 3-D object seen from different view angles

Motion features in a video sequence can provide useful clues regarding the temporal dimension of the video Four MPEG-7 motion descriptors have been defined to cover various aspects of motion

- *Motion activity* – This is defined as the pace of motion as perceived by viewers Typical examples of high activity include "car chasing" and "track racing" while examples of low action are "news anchor shot" and "an interview scene" In addition to the activity intensity, the descriptor considers three other attributes dominant direction, spatial and temporal distribution of activity The identified dominant activity direction can be assigned to any of eight equally spaced directions Spatial activity distribution indicates the number and size of active regions in a frame For instance, a news anchor shot would have one large active region, while an aerial view of a city street would have many small active regions Temporal activity distribution shows whether the activity is prolonged over the duration of a video segment or is restricted to only a part of the duration

- *Camera motion* – It is defined as the movement of the camera (or of the virtual viewpoint) in a video sequence The descriptor provides the following basic camera operations fixed, panning (horizontal rotation), tracking (horizontal transverse movement), tilting (vertical rotation), booming (vertical transverse movement), zooming (change of the focal length), dollying (translation along the optical axis) and rolling (rotation around the optical axis)

- *Motion trajectory* – It is defined as a spatio-temporal localisation of the centroid of a moving region The trajectory of the moving region consists of a set of keypoints and each keypoint records the spatio-temporal position of the region at each time interval

- Parametric motion – this is the evolution of arbitrary regions over time in terms of a 2-D geometric function The motion is characterised with reference to the panorama/mosaics of a video segment and it expresses motions that are not described by motion trajectory such as rotations or deformations

Finally, the MPEG-7 face-recognition descriptor is defined as the projection of an original face vector onto a space which consists of a set of basis vectors (i e possible face vectors)

Figure 2-2 below gives an example of the MPEG-7 edge-histogram descriptor that contains an element named BinCounts MPEG-7 defines the total number of histogram bins as 80 and the BinCounts element is an integer list with a length of 80 Each histogram bin is first given as the percentage of the total number of image blocks with the corresponding edge type and the total number of image blocks in each region It is then non-linearly quantised into an integer value between 0 and 7 based on the quantisation tables in [Cieplinski et al, 2001]

```
<EdgeHistogram>
        <BinCounts>
                2 4 5 3 4
                0 5 3 2 1
                2 4 7 2 6
                <!-- etc -->
        </BinCounts>
</EdgeHistogram>
```

Figure 2-2 An example of the MPEG-7 edge-histogram descriptor

Figure 2-3 below shows an example of the MPEG-7 scalable-colour descriptor Attribute *numberOfCoefficients* specifies the number of coefficients obtained from the Haar transformation in the representation The possible number of coefficients is

16, 32, 64, 128 and 256 The descriptor is scalable in terms of numbers of bins by varying the number of coefficients used Element *Coefficients* is described as a signed integer vector

```
<ScalableColor numberOfCoefficients = "16">
        <Coefficients>
                2  34  25  13
                8  19  21  10
                12  4   7  22
                <!-- etc -->
        </ Coefficients >
</ScalableColor>
```

Figure 2-3 An example of the MPEG-7 scalable-colour descriptor

The challenge for developing MPEG-7 visual descriptors is that they must be meaningful in the context of different applications The descriptors were chosen in extensive tests and their various characteristics were assessed before being adopted in the final standard [Manjunath et al, 2001] [Sikora, 2001] The compactness of the descriptor is achievable since the reconstruction of the original visual feature from it is not required

If only one MPEG-7 visual descriptor is available, search engines can be used to search, filter or browse the visual content according to a suitable similarity measures Most visual descriptors use the L1 norm, while some rely on the L2 norm and others adopt statistical distance measures If many MPEG-7 visual descriptors are presented, content matching for the descriptors are often implemented based on a weighted combination of visual descriptors and possibly including text features

Given a number of visual descriptors that are confined to a frame or group of frames, the next task left is how to combine them to depict a bigger picture in which the structural properties of a video are under consideration (i e temporal, spatial and spatio-temporal) MPEG-7 defines Description Schemes (DSs) for describing such combinations in a suitable way and we will give the details of the structural properties of the video in section 2 4 DSs is a description language about the organisation/appearance of descriptors in the final output, but does not provide a solution to the problem of content matching between images or videos for the

descriptors Our work attempts to solve the retrieval problem in which more than one visual descriptor along with text features is included in the structural representations of videos

2.4 The Structural Representations of Videos

The structural representation of a video focuses on segments that consists of *spatial*, *temporal* and *spatio-temporal* components A *spatial* segment is defined as a group of connected pixels such as a region in a still image A *temporal* segment is regarded as a sequence of continuous video frames, for instance, a shot A *spatio-temporal* segment is a combination of the spatial segment and temporal segment which is frequently used in object movement tracking

To study the temporal features of a video we need to decompose the video into segments and sub-segments, represented by a VideoSegment DS, or more specifically, into scenes and shots, annotated by a Shot DS Temporal subdivisions may allow gaps and overlaps between segments as illustrated in the bright blue sub-segments in Figure 2-4 It is also possible for a child segment to have multiple parent segments by referencing the child segment from the multiple parent segments using the SegmentRef DS Each subdivision may be further represented by one or more frames extracted from the subdivision A single frame extracted from a video sequence can also be considered a video segment in MPEG-7 A video segment may be described by visual features (colour, texture, shape and motion) and some elementary text information

The elementary text information could be any annotated text descriptions of a segment such as concept labels, or transcripts from a video or a TV programme The transcripts can be obtained via four different sources Our experiments for TRECVID2002 and TRECVID2003 use the text information created from ASR software (see Chapters 6 and 7)

- Subtitle/teletext signal from TV cable which can sometimes contain noises due to the quality of reception and transmission.
- Automatic Speech Recognition (ASR) outputs which also have degrees of noise due to the error rate of a ASR software.
- Subtitle information from a DVD which can suffer noise from the OCR process (since textual information is stored as images).
- Transcripts of some TV programmes made available on the web such as those for the most recent episodes of Horizon (BBC's science documentary series) since autumn 1999[6].

Figure 2-4 below illustrates an example of an MPEG-7 temporal structure representation. The recursive decomposition of a video can create a tree hierarchy of video segments. At the top level we have video segment *Seg0* with its media time stamp which contains the start time and duration of the segment in a video. *Seg0* is further divided into sub-segments based on the temporal decomposition and one of the sub-segments has an id *vs001*. Sub-segment *vs001* has three properties: (1) the media time stamp, (2) text information about the video segment, (3) one or more representative frames.

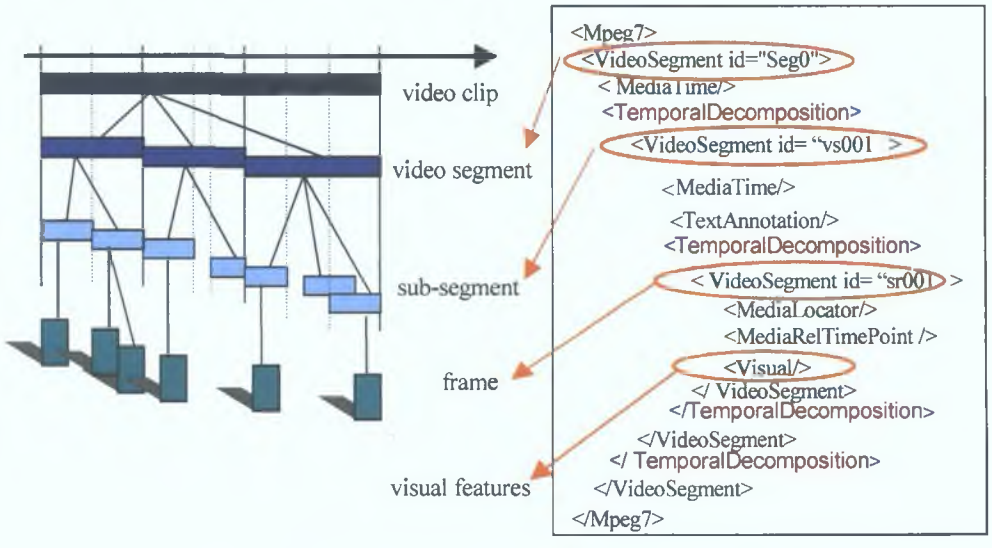


Figure 2-4: MPEG-7 temporal structure representation

[6] Horizon, BBC Science Documentary Series, started in 1996. The archive is available at http://www.bbc.co.uk/science/horizon/. Last visit on 10 March 2004.

A representative frame is described as a video segment in MPEG-7 such as *sr001* The representative frame is often extracted and physically stored in a different location from the video material for fast access purpose and a media locator is thus given Only a time point relative to the beginning of a video is used as the time stamp for the frame Various visual features as mentioned in section 2 3 can be embedded in the representative frame using an element called *Visual*

2.5 Related Work on MPEG-7 Searching

Over the past decade, multimedia content has been steadily migrated from the original analogue mediums (VHS tapes for example) to digital forms and we can expect this to continue With more and more image and video becoming available on the Web, indexing and annotation tools will be required to catalogue and facilitate search over these files With the introduction of MPEG-7, a description of a multimedia file can be generated and used by search engines New generations of search engines will be required to interpret the image or video clips that we are interested in and compare it to all the files on the Web, much as we do today for Web pages

But Web pages are very different from multimedia files When automatically indexing an image file, we index differently than we would when indexing a Web page of which the statistical properties of terms are used For example, we extract primitive visual features such as colour, texture and shape from images/videos and probably attach text information to them if appropriate annotation tools are available and there are enough human resources assigned for the task

Much current research in MPEG-7 searching takes on two main approaches (1) visual features only, (2) text features and the MPEG-7 structure The first approach makes use of visual features such as colour, texture and shape or any combinations and results in variants of image search engines Similarity measures such as L1 or L2 norm are widely adopted for this purpose In dealing with large amounts of image/video content, a two-stage image (or video shot) selection process is often employed in order to speed up the retrieval process Given multiple visual descriptors, each descriptor can be used separately (in parallel) for finding a set of

image candidates The created subsets are combined into a larger set for the final selection process and more computational demanding comparison methods for the rank output is then employed

For instance, Ngo et al implement a two-level hierarchical K-means clustering approach to group shots with similar colour and motion using a top-down approach, where the colour is utilised at the top level with motion at the bottom level [Ngo et al, 2001]

A cluster centroid at the top level represents the colour properties of a cluster while a cluster centroid at the bottom level describes the motion properties of a cluster During retrieval, we first compare cluster centroids at the top level with the colour properties of a query A cluster with the nearest centroid is selected and its sub-clusters at the bottom level are further compared with the query Each sub-cluster is orderly considered based on its nearest to the motion properties of the query and the members in the sub-cluster are ranked in ascending order of their distance to the query After all sub-clusters are sorted, the next similar cluster at the top level is handled in the same way This process is repeated until the most similar clusters at the top level are visited

The second approach for MPEG-7 searching involves indexing text features and possibly the MPEG-7 structure The two following methods dedicated to MPEG-7 retrieval using text features are explained the SAMBITS project and the MPEG-7 based inference network

The SAMBITS project has developed an MPEG-7 terminal built upon the HySpirit (Hypermedia System with Probabilistic Inference for the Retrieval of Information) XML search engine [Lalmas et al, 2001] [Pearmain et al, 2002] HySpirit combines probabilistic inference with *4-value logic*, allowing for structured document retrieval and possible inconsistencies between the content of document nodes In contrast to a *2-value logic* in which an interpretation of a fact or event only contains two values true and false, *4-value logic* defines four values true (T), false (F), inconsistent (I) and unknown (U), and the corresponding truth tables is given in Table 2-1 [Fuhr & Rolleke, 1998]

Table 2-1 The truth tables for 4-value logic

AND	T	F	U	I
T	T	F	U	I
F	F	F	F	F
U	U	F	U	F
I	I	F	F	I

OR	T	F	U	I
T	T	T	T	T
F	T	F	U	I
U	T	U	U	T
I	T	I	T	I

NOT	T	F	U	I
	F	T	U	I

The logical inconsistencies of document content are present when several parts of a document or from different documents (i e hyperlink) are combined to form an answer to a query For instance, a document *doc1* contains two sections *sA* and *sB* Section *sA* have terms *furniture* and *landscape* and the corresponding logical formula is *furniture* Λ *landscape*, while section *sB* is more about *woodwork* but not *landscape* and the formula is formed as *woodwork* $\Lambda \neg$ *landscape* Given a query looking for *woodwork* Λ *landscape*, a set {*sA, sB*}, namely *doc1*, is not an answer since inconsistency exists with respect to proposition *landscape* where the combined formula is *furniture* Λ *landscape* Λ *woodwork* $\Lambda \neg$ *landscape*

In 2-value logic, the set {*sA, sB*} would be the answer to *furniture* Λ *woodwork* Λ *landscape*, or *furniture* Λ *woodwork* $\Lambda \neg$ *landscape* In 4-value logic, the set {*sA, sB*} is an answer to *furniture* Λ *woodwork* only

The HySpirit search engine specifies the probability of an event being true and false but also being inconsistent in a triple value form as *true/false/inconsistent* The probability for the unknown event can be obtained as the complement to 1 The estimation of the probability that a document implies a query considers not only the present and absence of terms, but also the inconsistency of terms if parts of document or different documents are taken into account

An MPEG-7 retrieval model built upon HySpirit uses the standard Term Frequency and Inverse Document Frequency (TF*IDF) as the relevance ranking function, considers the temporal and spatial relationships between segments, and supports the inheritance of segments that have a direct parent-children relationship Queries to this model are text-based, involving content (e g keywords) and/or meta-data (e g authors and titles) queries, possibly with an XML structure The model integrates keyword-based, fact-based and structural information for MPEG-7 retrieval

The second MPEG-7 retrieval model using only text features has been designed based on inference networks and was developed to capture the structural, content-based and positional information by a Document and Query network [Graves & Lalmas, 2002] Content-based information is text-based Positional information is context-related and contains the location of content within a document Based on Turtle and Croft's inference network model for text retrieval [Turtle & Croft, 1991], the MPEG-7 inference network for a document has three layers Document, Contextual and Conceptual

Two types of conditional probability in the *contextual* layer are distinguished (1) the *structural probability* estimated by the duration ratio between a video segment and its parent segment, (2) the *contextual probability* determined by the sibling information – the inverse of the total number of siblings The *structural probability* is defined based on the temporal characteristic of videos while the *contextual probability* is thought to be the attributes of a video segment such as "Video_Segment -> Creation_Information" and "Video_Segment -> Media_Information"

The *conceptual* layer consists of content nodes that represent the content of the MPEG-7 document nodes such as terms identified in the Text_Annotation DS or in the Title DS The conditional probability between a context and a concept is defined based on the Term Frequency (TF) and Inverse Document Frequency (IDF) of the concept

The Query network consists of concept and context nodes and query operators as described in the Document network Retrieval is done in a bottom-up fashion by accumulating evidence (i e probability) from all layers to find the best shots

The two MPEG-7 retrieval systems described previously use only text features (i e metadata and manual text annotation) at the shot level and take into account the structural characteristics of video The major difference is that the HySpirit based method employs probabilistic 4-value logic to include inconsistency events during the access from a parent context to its sub-contexts or siblings, while Graves & Lalmas' method based on document inference network uses probabilistic 2-value logic and introduces different types of conditional probability to estimate the weights of all accessible nodes

However what they both lack is a means of integrating visual descriptors into the models Although automatic assignment of textual features to images/video shots has been developed using the text from closed-captions/subtitles and transcripts to greatly reduce the labour involved in manual assignment, many images are still without associated text information A user's image need may occur at a primitive level that maps immediately onto the visual features of an image These features can be best represented by image examples and retrieved by systems utilising pattern matching techniques based on colour, texture, shape and other visual features Thus we come to using visual features for MPEG-7 searching, as mentioned in the beginning of this section, but these approaches are also restricted since no textual and structural features of videos are included

A video indexing approach is hybrid in nature since a user's need can be made in two different ways keywords and image examples (or key frames extracted from a video) But a need composed in either way has its limitations When asking for shots related to some given search keywords, the difficulty of locating the relevant shots may occur if no relevant text is attached to the shots When asking for shots similar to a given photo, the difficulty of finding the relevant shots is due to the existence of a gap between the primitive visual features and the higher level cognition It seems necessary to tag these features into terms (or concept labels) that occur to users in the course of a search But a degree of misjudgement appears to be

inherent in all visual-based indexing approaches A photo of a sunset by the sea, for example, is characterised with large amount of redness and a search with this photo would also retrieve shots of a sunrise or even a fire scene

Our work takes advantage of the hybrid nature of the video indexing approach, namely to use both visual and textual features in order to correct the misjudgements in video retrieval We embed key frame information as a low level in the MPEG-7 hierarchical structure including the colour and texture descriptors, as shown in Figure 2-4 We propose to employ K-means shot clustering algorithm to organise shots of each video at the primitive visual feature level If isomorphic relationships can be expected to occur among the shots, it is reasonable to assume that terms collected from transcripts can be linked to shots by their proximity to the corresponding cluster centroids and that a query image should suggest query terms by its proximity to cluster centroids in the collection

The mapping between the visual features and terms serves two purposes (1) to narrow the semantic gap between them, (2) to provide a way to combine the visual and text features of shots Having obtained the mapping, we discard the visual features and use the derived text resulted from the mapping in the course of indexing We suggest that the meaning of shots can be enriched by aggregating both the original and derived text and this would be useful in cases where shots have no accompanying text This approach will be introduced in Chapter 4 and its effectiveness will be tested as an automatic search defined in the TRECVID experiments (see Chapter 6 and 7)

2.6 Conclusions

The MPEG-7 standard is a set of description structures with associated meanings defined and recognised in public domains Since the goal of MPEG-7 is to maximise interoperability, the scope and possible application domains covered are very broad such as retrieval, navigation and user interaction MPEG-7 is only a multimedia representation standard and defines nothing about the implementation of any applications An application will not need the complete set of MPEG-7 descriptions

but only certain parts according to the specific requirements of the application. The implementation issues associated with extracting features from video which can be encoded in MPEG-7 are left to individual research groups.

For a video retrieval application, we are often concerned with the temporal structure and various features that can be computationally acquired from videos, all of which have corresponding descriptions in the current MPEG-7 standard. We have seen two retrieval methods specifically designed for MPEG-7 documents both of which use probabilistic inference to solve the problem with the retrieval of structure documents. An MPEG-7 description is equivalent to a XML document if only text features are considered. The XML retrieval approaches available to date study the statistical properties of terms and the hierarchical structure of documents. We will continue to review the other three methods for searching XML documents in next chapter, namely information retrieval based, path expression and tree matching.

XML-based access to images or videos has its limitation since many images and videos are not tagged to the appropriate description such as keywords and concept labels. As a result increasing interest in the development of image-based (content-based) solution has been prompted such as colour, texture and shape. Image-based access to MPEG-7 visual feature descriptors is different from the XML-based approaches and it relies on the characterisation and pattern matching of primitive visual features. But image-based searches are restricted by image examples and can often produce unexpected results.

We have seen the need to explore the combined use of both text-based and content-based techniques for MPEG-7 searching. The interest of this work focuses on a method to integrate visual information into an available XML retrieval approach and our approach will be given in Chapter 4. We propose to employ K-means shot clustering algorithm to organise shots of each video at the visual feature level. If isomorphic relationships can be expected to occur among the shots, it is reasonable to assume that shots within the same cluster are similar visually but also semantically to some extent. Terms collected from the transcripts of member shots in a cluster can be linked to the shots by their proximity to the corresponding cluster

centroid We suggest that the meaning of a shot can be enriched by including the text derived from its associated cluster into the original text

Concerning structured document retrieval for MPEG-7, coping with document structure is not a difficult task since the query formulation has to refer explicitly to the document structure XML document retrieval allows for different ways of abstract representations from document structure and any nodes in the hierarchy can be retrieved The two approaches for MPEG-7 searching (i e HySpirit-based and Graves & Lalmas' method) consider the statistical properties of nodes in a bottom-up fashion The weights of all leaf nodes are calculated and they are then used in computing the weights of their parent nodes The process is repeated until the root is reached

The retrieval unit defined in our TRECVID experiments for video retrieval is restricted to a shot and the shot is in the middle level of the MPEG-7 hierarchy One level below a shot is a key frame and one level above is a scene or a story (if either scene or story can be determined), and one level above the scene/story is an entire video Our work focuses on the estimation of weights at the key frame level and at the shot level and we skip the nodes at the higher level

Chapter Three

A Review of Approaches to XML Document Retrieval

An MPEG-7 description of a multimedia document like a video can be regarded as a standardised XML document if only text features of that document are under consideration In this chapter we extend our review of current XML document retrieval approaches in order to address their strengths and potential in terms of video retrieval We classify the techniques for XML document retrieval into three categories (1) Information Retrieval based (i e IR-based), (2) Path Expression and (3) Tree Matching Each approach is discussed in turn We suggest that IR-based approaches provide a better foundation for MPEG-7 video retrieval since they allow for content matching and approximate structure matching, as described in section 3 4

In section 3 5, we explain the aggregation technique used in an IR-based XML document retrieval approach Aggregation was originally introduced as a query language model by Yager [Yager, 2000] and it provides a way to combine scores from various aspects of a query according to associated importance values The importance value is a measurable quantity of a user's degree of satisfaction The satisfaction degree is usually reflected in natural language words such as "mostly" and "a little" Each word has a corresponding function to specify importance values The word and function pair is defined as a linguistic quantifier, which will be described in section 3 5 2

3.1 An Information Retrieval based Approach to XML Document Retrieval

The Information Retrieval approach considers both the structure and the content of documents The TF*IDF vector space model and its variants is usually used to index the document content Returned items are ranked based on the dot products between the index and a given query Two different methods of using the document structure are under consideration passage retrieval and aggregation-based retrieval

Passage retrieval applies the conventional TF*IDF to index a whole document as well as its children sections and subsections *TF* gives the frequency of a term t_j occurring in a document *IDF* is the inverse of the total number of documents containing term t_j If the content of documents is being compared, the indexing unit is a document and document frequency is defined as the number of documents containing the relevant term If the content of sections is being compared, the indexing unit is a section and document frequency is determined by the number of sections containing the relevant term During retrieval, items returned for users could be documents, sections or subsections sorted in ranked order Experiments have shown that the content and structure of documents is useful for text retrieval High precision can be achieved by combining information from both the whole document and their sections but no gains for high recall retrieval [Wilkinson, 1994]

Aggregation-based retrieval treats documents and their children as contexts It computes the overall weight of a term within a context c by aggregating the weight of its own content and those of its structurally related contexts, namely its children but also siblings and referenced contexts [Kazai et al, 2001a] [Kazai et al, 2001b] Each context is associated with an aggregated term weight vector Given a keyword query, the most relevant contexts are located as the final ranked results by the dot products between the term weight vectors of all contexts and the query vector For instance, if context c contains three sub-contexts s_i ($1 \leq i \leq 3$) and each sub-context has a term with weight t_i We associate an importance value w_i with each sub-context and obtain the overall weight t_c in context c according to $t_c = \Sigma_i w_i t_i (0 \leq i \leq 3)$, where t_0 is the term weight in context c with an importance value w_0

The importance value is a measurable quantity of users' degree of satisfaction and is specified indirectly by the users who may find it easier to express their satisfaction in natural language words such as "most" and "at least" Each natural language word has a corresponding function to calculate the importance value When choosing the word "at least", a user would like to put most of the important values on the higher term weights to emphasise the higher scores, or choosing the word "mostly", to give most importance on the lower term weights to emphasise the lower scores, or choosing word "some" to get even importance for the average The details of using an aggregation approach for structured documents is given in section 3 5

Traditional information retrieval approaches can be directly applied to structured document retrieval since structured documents are text documents with additional mark-up Passage retrieval techniques were originally developed to solve the problem with the retrieval of long documents in which a user might have difficulty to extract the relevant parts to the query [Wilkinson, 1994]

Both passage-based and aggregation-based approaches take advantage of the important contextual information of structured documents They differ in the way of obtaining the weight of a term in a given context or section The passage-based approach takes into account of contextual information in the TF*IDF formula where the indexing units are different accordingly whereas the aggregation-based approach deals with it by aggregating the TF*IDF weights of terms in a context and its structurally related contexts

Both approaches take in informally phrased keywords as part of a query and return the most relevant contexts to the query As a result users do not need to post a query with XML structures, for example "find a particular context c which is particularly about w", which is an advantage when users do not know the document structures in a collection

3.2 The Path Expression Approach to XML Document Retrieval

The Path Expression approach to XML document retrieval mainly studies the XML document structure rather than the content It regards an XML document D as a root ordered graph The graph contains a set of nodes (i e element ID) $N = \{v_0, v_1, \quad v_N\}$ and a set of edges $E = \{ e_0, e_1, \quad e_N \}$ v_0 is the root node of the document Each edge is labeled with either an XML element name or attribute name A label path is defined as a path from root v_0 to any given node $v \in \{v_1, \quad v_N\}$ $LP_v = \{ v_0, e_1, v_1, e_2$ $e_v, v\}$

A path index is constructed by looking up each node $v \in \{v_1, \quad v_N\}$ in the graph, putting nodes with the same label path in a group, and assigning the label path as a search key to the group Other information can also be attached to the group such as term frequency and other nodes that occur in the path The path index can become large when the label paths overlap (where the label path of node v is included in its children's label path) For example, let $v, w \in N$, node v is the parent of node w if and only if label path LP_v is included in label path LP_w, that is $LP_w = LP_v \cup \{e_{vw}, w\}$ $= \{ v_0, e_1, v_1 \quad e_v, v, e_w, \omega \}$

A number of methods have been proposed to address the label path overlap problem, such as Dataguides [McHugh et al, 1997] and ToXin [Rizzolo & Mendelzon, 2001] We will use a data graph corresponding to a simple XML document for illustration (Figure 3-1)

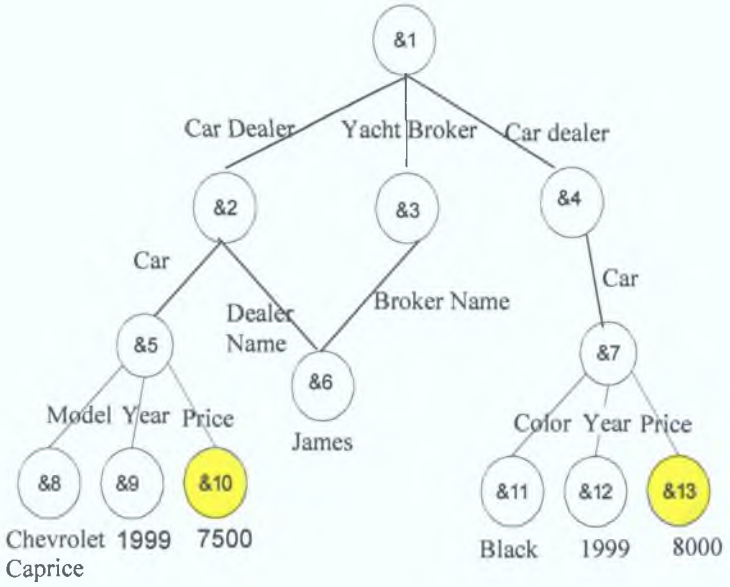


Figure 3-1: A fragment of a simple XML document in a data graph

A Dataguide, is primarily developed for schema-free environments, maintaining the structural summaries of an XML collection by redefining the label path of a node v as LP_v={ e_1, e_2 e_v}. An instance of label path is called a data path DP_v and is denoted as DP_v={ e_1, v_1, e_2, v_2 e_j, v }.

A Dataguide is built by removing all text nodes (i.e. leaves) in the graph since they make no contribution to the structure summary. Each possible label path is recorded only once in the remaining nodes (i.e. non-leaves). Each label path is then attached to a target set of element IDs that are the last element ID from some data path instances which satisfy a given label path. The Dataguide structure, also called the path index, can help answer queries like "locate all elements reachable via a given label path" by simply looking up the label path index to return the corresponding target set.

For instance, the data path DP: "Car_dealer.&22.Car.&24.Price.&29 is the only instance of label path LP: "Car_dealer.Car.Price" in Figure3-2. The target set of "Car_dealer.Car.Price" is {&10, &13} where both &10 and &13 are the last element IDs that have a label path "Car_dealer.Car.Price" in the source XML files.

In dataguides, all text nodes are indexed in two different ways depending on the types of the text content: (1) value index such as simple string, real data and date, (2) text index which contains term frequency information.

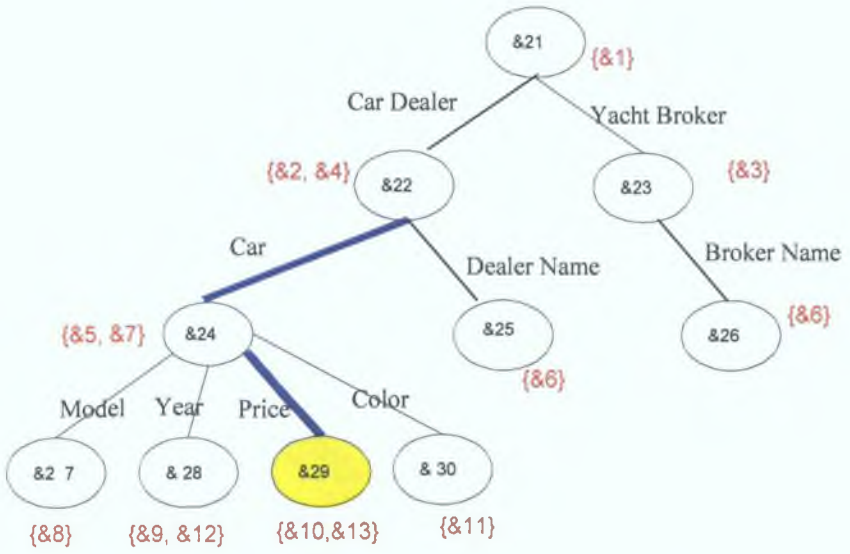


Figure 3-2: A Dataguide to the sample XML fragment

Dataguides are useful for navigating a path from the root but lack facilities for locating an arbitrary node and backward navigation. A Link index structure is then introduced to support an inverse child-parent relationship look-up [McHugh et al, 1998]. For instance, given a child element "&29 and a label "price", return all parents {&24} such that there is a "price" – labelled edge from &24 to &29.

ToXin [Rizzolo & Mendelzon, 2001] builds on dataguides by accommodating both forward and backward navigation. Two types of index are used: the path index and the value index. The path index consists of two components: the index tree which is equivalent to a Dataguide (see Figure 3-3) and a set of instance functions. Each edge in the index tree has an instance function to keep records of the parent-child relationship between the pairs of nodes that are linked by the edge in the source XML documents. Each instance function is stored in two redundant hash tables: a forward instance table keyed on parent nodes (see Figure 3-4) and a backward instance table keyed on child nodes. All leaf nodes are stored as the value index and keyed on parent nodes (see Figure 3-4).

Dataguides and ToXin differ in the way of keeping the parent-child relationship table. In a Dataguide, the instances of a label path are stored in its target set, in which no information about the parent-child relationships is kept. For example, the target set of "Car_dealer.Car" is {&5, &7} but no information is available that &2 and &4 are the parents, respectively (see Figure 3-2). A separated look-up table is maintained in Dataguides to solve the backward problem. In a forward instance index of ToXin, nodes &5 and &7 are the instances of the path "Car_dealer.Car" and have parents &2 and &4 via reference table IT2 (see Figure 3-4).

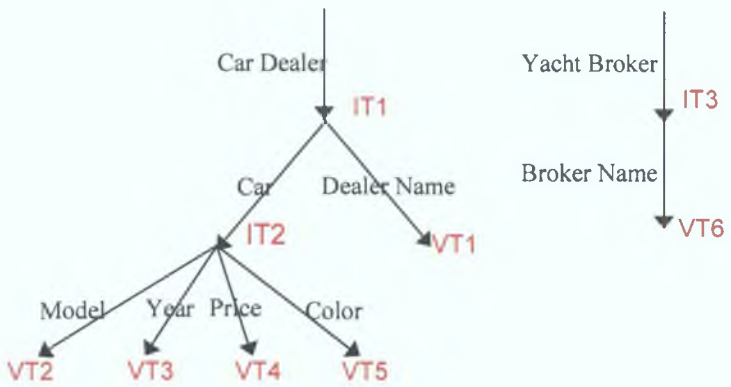


Figure 3-3: An index tree of ToXin

IT1

Parent	Child
1	2
1	4

IT2

Parent	Child
2	5
4	7

IT3

Parent	Child
1	3

VT1

Node	Value
2	James

VT2

Node	Value
5	Chevrolet Caprice

VT3

Node	Value
5	1999
7	1999

VT4

Node	Value
5	7500
7	8000

VT5

Node	Value
7	black

VT6

Node	Value
3	James

Figure 3-4: A forward instance index and value index of ToXin

The path expression approaches search for particular words in particular mark-up tag appearance and items returned are restricted to elements satisfying the query structure. Unlike the IR-based approaches in which the content index plays an

important role, the path expression approaches index the content and structure of documents in two different forms The structure of all documents is stored in a summary table with pointers to the source XML parts Although content search is supported via a value index table, structural search is the main thrust of the approach

3.3 The Tree Matching Approach to XML Document Retrieval

Sometimes XML queries can be expressed by a tree pattern and searching can be converted to a tree pattern-matching problem, to determine whether the query tree exactly or approximately appears in the XML data tree Figure 3-5 shows a sample XML query tree, looking for car models which are priced below 7000 and a possible XML fragment to the query is found and highlighted in grey

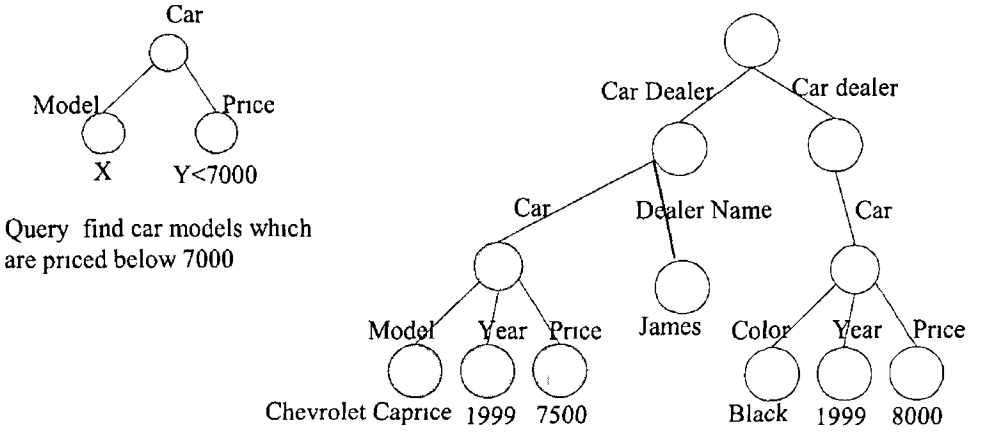


Figure 3-5 A sample XML query tree

Exact tree pattern matching determines the distance between two given trees as measured by the minimum cost of edit operations that are required to transform one tree into the other (e g substitution, insertion and deletion) [Tai, 1979] *Approximate* tree pattern matching is measured by the number of paths in the query tree that match against the data tree [Chen et al, 2001] or other distance functions

[Schlieder & Naumann, 2000] [Sheridan & Smeaton, 1992], which are described as below

An approximate tree matching approach estimates the number of twig matches of a query tree in a data tree [Chen et al, 2001] A twig is a small fragment of a source tree A data tree is decomposed into a number of twigs and only frequently occurring twigs are kept in a summary table due to the large number of possible twigs The frequency information for each twig in the data tree is estimated and stored in the summary table Given a tree query, the number of matches is estimated by first dividing the query into a set of twigs based on the available twigs in the summary table, then estimating the number of matches of each query twig in a data tree, finally combining the scores of all query twigs into one final score for the entire data tree

Another method of handling an XML query tree is to use approximate embedding tree techniques [Schlieder & Naumann, 2000] An embedding data tree is retrieved when the data tree approximately embeds the query tree such that the labels and the predecessor-successor relationship of the nodes are preserved The quality of an embedding is measured by a deletion cost function An embedding will have a low cost and be ranked high when a small number of nodes are skipped to embed the query tree A high cost indicates that large differences exist between the query and the data sub-tree

A given query tree is processed in a bottom-up manner For each query node q, a query sub-tree rooted at q is created and the sub-tree contains the child nodes of q A set of valid embeddings of the query sub-tree in data sub-trees can be found and only one embedding that has the minimal cost is kept The process is repeated for each node in the query tree and a set of embeddings is then formed and sorted by increasing cost for users

XML documents can be represented in a tree hierarchy The motivation for introducing tree matching is to provide a means of locating fragments in a target tree by giving a pattern tree that describes the structural occurrences requested by users However, a problem arises because users may not know what kind of query

specification is the most appropriate since the document structures are unknown to them

In [Sheridan & Smeaton, 1992], an approximate tree matching technique was described to assist the matching between phrases for a Natural Language Processing (NLP) task, called the Tree Structured Analytics (TSAs) NLP queries and NLP documents are each reduced to trees, where the trees are derived from syntactic parsing of the phrases Matching is conducted using the reverse post-order tree-traversal strategy Content (words) matching at the lowest level is identified based on the text form of the word, its morphological base form, and its syntactic function label Structure matching at the higher level is then examined based on the syntactic relationship Different scores are assigned to the different types of matching (e g identical syntax match or base form match) All individual match scores are added together to give the overall score for the entire TSA structure

The tree matching approaches to XML document retrieval are more concerned with the structure than the content of structured documents because frequency information of the structure is carefully estimated A text search with specified XML structure is supported and performed in two stages Structural search is first carried out to find the appropriate fragments in target trees, and these fragments are further examined and ranked based on other criteria such as proximity content search But TSAs matching is carried out in a reverse way by firstly finding content matches before structure matching takes place

3.4 Summary of XML Document Retrieval Approaches

Three types of XML indexing approaches are compared and we summarised in Table 3-1 based on four criteria (1) XML schema constraint, (2) structure matching, (3) content matching and (4) query formulation

In the Information Retrieval based approach, searching is performed based on the similarity between the content of documents and a given query Structural information is included in the content representations before the final content index

is created No extra cost is spent on maintaining the document structures, therefore, simple but not complicated document structure matching is indirectly supported

In particular, the passage retrieval approach is designed to deal in a schema-constraint environment since the weighting formula is conditioned upon the pre-defined indexing units and these units could be varied (e g a document, section or subsection) On the contrary, the aggregation-based approach can be used for schema-free XML documents because the underlying weighting formula remains unchanged and document structural information is considered as additional evidence to the content

The IR-based approach sorts results based on the content similarity regardless of the document structure and they could be either whole documents or subsections If users are not interested in strictly structured results but more concerned with the returned relevant items, the approach is adequate Also this approach is useful when the document structures are unknown to users and relevant information can be found based on full text search alone

In the Path Expression approach, the document content and structural information are indexed separately A value index is used to store the statistical properties of the content information and a path index is created to summarise the structure of all documents in the collection Each path has a pointer to a set of nodes in the source XML documents that satisfy the path expression During retrieval, we look up the path index for a user's requested path, return a set of nodes pointed by the path and sort them in order The approach needs to contain detailed information about the structures in order to deal with the exact structure matching, thus requiring a relatively large amount of effort on maintaining the different indexes Also required in this approach is the query conversion from a text description to a path expression

In the Tree Matching approach, searching is performed based on the similarity between fragments of documents and those of a given query tree Results are sorted based on the statistical properties of document structures rather than document content It supports complex structure matching including exact matching and

approximate matching Storage and computational overhead is thus paid for this complexity An initial search based on structure information is often followed by a simple text search Full text search alone is not under consideration in this approach Another problem with this approach is that defining document structures in a query is necessary and the query is required to convert to a tree pattern

Table 3-1 A summary of the three XML document retrieval approaches

	Information-retrieval based	**Path Expression**	**Tree Matching**
XML schema constraint	Schema-constraint XML documents if a weighting formula takes into account document structure, schema-free XML documents otherwise	No, schema-free XML documents	No, schema-free XML documents
Structure matching	Approximate structure matching No rank is provided based on the XML structure	Exact structure matching Results are mostly based on XML path expressions All items pointed by the path are considered for content ranking if a given query path exists in the path index	Exact & approximate structure matching Results are sorted based on the scores of the XML structure rather than content
Content matching	Yes, value index is provided Results are sorted based on the rank regardless of XML structure Results could be either full documents or subsections	Yes, value index is provided Ranking can be generated based on the content if the associated path information is qualified	No No value index provided
Query formulation	Mostly based on search terms	Based on XML structure conditioned on search terms or values	Based on a query tree pattern, requires converting a text query into a query tree

The requirements for video retrieval of MPEG-7 descriptor documents are now considered based on the four criteria previously discussed MPEG-7 documents are schema-constrained XML in general since they strictly conform to the schema definition However, they can also be seen as schema-free XML documents due to the broad coverage of MPEG-7 description tools An MPEG-7 document does not contain all the available description tools but in fact only a few selected description tools based on the requirements of the application domain For instance, an MPEG-7

document may describe the shot information of a video or may also contain key frame information of each shot concerning primitive visual features

Exact structure matching studies a restricted form of unification between the source data and a query Much effort including storage and computational requirements would be required in order to find good structures Approximate structure matching is introduced to reduce this burden so as to locate similar occurrences of a structure pattern (i e structure mismatches, deletions and insertions are allowed) The results found by approximate matching can be thought of as the superset of those found by exact matching The "best" approximate structures are ranked by their relevant scores to the query structure In video retrieval, approximate structure matching would be more suitable because users may not know or need to know the hidden video structures designed for indexing Moreover, in the context of large video collections, a shot is often defined as the retrieval unit and provides the accessible point to videos so that users can start browsing the video for further examination

In video content matching, a shot is considered to be relevant to a given query if it contains the minimal content that is relevant to the query Again content matching is approximate and a content rankmg/weighting method is required Two types of content are of concern in video retrieval text and visual features A chosen XML retrieval model is needed to facilitate both types of content matching This may be a difficult choice since none of the three types of approach are designed to accommodate the matching of visual features

Finally, a user's query to a video collection often has three different forms text, image and video clip example A query based on video structure is seldom the case since users may not know which of the retrieval units are the most suitable to represent their information need, whether a shot, a story or a scene A shot is therefore defined as the default retrieval unit to provide accessible points to videos in most video retrieval system

In short, a video retrieval model for MPEG-7 is required to deal with content matching of text and visual features possibly in a schema-free environment We chose the aggregation-based approach as the foundation for MPEG-7 retrieval since it satisfies most of the listed criteria and the matching of visual features can be easily adapted to the model The aggregation-based approach assumes that document structural information is treated as additional evidence to document content

Following a similar idea, we assume that visual features are another way of obtaining auxiliary evidence to document content We continue to consider that the visual features of each shot can be assigned to a term vector whose value expresses the shot's additional semantic content Before moving onto our video retrieval approach for MPEG-7 in Chapter 4, we further study the aggregation-based approach for structured document retrieval in the next section of this Chapter

3.5 Aggregation of Term Weights

The original aggregation method evaluates a user's query by combining scores from various aspects according to an associated importance value [Yager, 2000] Instead of specifying the importance of an aspect explicitly in a numerical way, the aggregation method only requires the user to determine the importance using some natural language words and automatically maps the selected words onto a corresponding mathematical function to obtain the actual importance values The word and function pair is called a linguistic quantifier

For example, a document D is associated with the A_j ($1 \leq j \leq n$) concept and a score $A_j(D)$ indicates the extent to which document D is about concept A_j A query to a system can be constructed based on any AND/OR combination of these concepts such as (A_1 AND A_2 AND A_3) and the respective importance values of these concepts w_j, ((A_1,w_1) , (A_2,w_2) , (A_3,w_3)) Users generally do not give an exact value for each w_j and instead phrase their queries using natural language, for example "I want a document that is *mostly* about concept A_1 and contains *a little* of concept A_2"

Linguistic quantifiers are thus introduced to translate a users' degree of satisfaction into a measurable value w_j

Kazai et al further extended Yager's query language model into a retrieval method for structured documents [Kazai et al, 2001a] A structured document is represented as a graph whose nodes are elements within a document hierarchy such as document, chapter and paragraph, and whose edges reflect structural relationships between the connected nodes Three types of structural relationships are of interest hierarchical (parent-child), linear (siblings) and referential (hyperlink)

A document component C is now not only associated with its own content A_j^C but also the content A_j^{Si} of its structurally related components S_i ($1 \leq i \leq k$) The assumption of document components being about concepts is replaced by those containing terms that make a good concept as indicated by $t_j(A_j^C) = A_j^C\ (C)$ An overall score for C is obtained by combining the weights in content A_j^C and A_j^{Si} based on three factors the type of structural relationships, the type of content and the linguistic quantifiers used The aggregation model used in structured document retrieval is introduced below including OWA operators and linguistic quantifiers

3 5 1 Ordered Weighted Averaging (OWA) Operators in Structured Document Retrieval

Given a document node C, we have the weight t_0 of a term in its own content A_0^C and the weights t_k in its k structurally related contents A_j^c ($1 \leq j \leq k$) Vector $t = \{ t_0, t_1, \ldots, t_k \}$ is the argument vector of the term and vector $w = \{ w_0, w_1, \ldots, w_k \}$ indicates the importance value associated with the corresponding term weight of component An *Ordered Weighted Averaging* (OWA) operation with respect to t is obtained as follows

1 Sort entries of vector t in descending order and obtain the *ordered argument vector* $b = \{ b_0 , b_1 , \quad b_k \}$, where b_j is the j-th largest of t_i The ordering of the corresponding importance weighting vector w is changed to that of vector b, denoted as $\alpha = \{ \alpha_0 , \alpha_1 , \quad \alpha_k \}$ where α_j is the importance value of the j-th largest of t_i

2 Apply the OWA operator as follows to obtain the overall score t^C of component C

$$F(t) = \sum_{j=0}^{k} \alpha_j b_j \quad \text{where} \quad \alpha_j \in [0,1] \quad \text{and} \quad \sum_{j=0}^{k} \alpha_j = 1 \tag{3-1}$$

The final score t^C has as its bound Min $\{ t_j \} \leq t^C \leq$ Max $\{ t_j \}$ and reflects the contributions from different document components

3 5 2 Linguistic Quantifiers

Having obtained the argument vector t, the next task left in aggregation is to determine a suitable importance weighting vector w Linguistic expressions are a popular method used in structured document retrieval [Yager, 2000] Words such as *"all"*, *"most"* and *"at least one"* are called *linguistic quantifiers* describing a proportion of items in a collection Each quantifier is related to a regularly increasing monotonic function Q Examples of Q are $Q = r^p$ where p ranges $[1,\infty)$, when p = 1 the function implies quantifier *some*, when p = 2, *most*, when $p = \infty$, *all*

In XML document retrieval, it is possible to determine the ordered importance weighting vector α by some measurable property of the document such as term weights t_j instead of users' explicitly defined importance Each entry of α_j can be calculated based on the formula below instead of the general normalisation technique to achieve $\alpha_j \in [0,1]$ and $\Sigma_j \alpha_j = 1$

$$\alpha_j = Q(S_j) - Q(S_{j-1}) \quad \text{where} \quad S_j = \frac{\sum_{j=0}^{j} b_j}{T} \quad \text{and} \quad T = \sum_{j=0}^{k} b_j \tag{3-2}$$

T is the sum of all ordered argument weights b_j of a term and S_j is the sum of j ordered argument weights of the j-th most satisfied components The quantifier *most* defined by $Q = r^2$ places most of the importance on the lower weights b_j thereby emphasizing the lower scores In contrast the quantifier "*at least ½*" defined by $Q = r^{1/2}$ emphasizes the top of weight list b_j, which was later defined in [Kazai et al, 2001a] We use the quantifier "*at least ½*" defined by $Q = r^{1/2}$ to guide the aggregation in the rest of our work

3.6 Conclusions

In this Chapter we reviewed three main approaches to XML document retrieval (1) information retrieval based, (2) path expression based and (3) tree matching based The IR-based approach incorporates the properties of document structure into the weighting of document content and relevant information can be located based on full text search alone It is therefore useful when the document structures are unknown to users

The path expression approach handles document structure and content in different ways a path index is used for storing the summary of document structure and a value index for the properties of document content Each path in the structure summary table has a link to all instances (i e elements) in source XML documents that satisfy the path expression Full text search can be achieved by looking up the value index and structure search can be performed by the path index The approach is flexible since novice user may use it like a text-search system and expert users may benefit from the document structures thereby getting more precise answers Storage overhead however is paid for this flexibility since detailed information about the structure summary is needed A disadvantage of the approach is that approximate structure matching is not supported, which is the generalisation of exact matching to find similar (not just exact) occurrences of a pattern in source files

The tree matching approach focuses on the properties of document structures but not the content During indexing, a document is decomposed into small tree

fragments of which the statistical properties are stored During retrieval, a query is also broken down into fragments and a matching between a document and the query is estimated by the cost of edit operations or probability that are required to transform the document tree into the query tree Complex structure matching including exact matching and approximate matching is supported in the approach The major disadvantages are that (1) full text search alone is not the initial intention of the design of the approach and (2) defining document structures in a query is necessary and the query is required to convert to a tree pattern before retrieval

The TRECVID experiments are carried out within the framework of a collaborative evaluation of video information called TRECVID which we will describe later As the retrieval unit for video defined in our experiments is restricted to a shot in the TRECVID experiments to be reported later in this thesis, structure matching for MPEG-7 retrieval may not be necessary However, a video retrieval model for MPEG-7 is in fact required in order to deal with content matching of text along with the retrieval using visual features

We chose one of the IR-based approaches, namely the aggregation-based method, as the foundation for MPEG-7 retrieval in the work in this thesis since structural information is incorporated into the weighting of content of document elements and the matching of visual features can be easily adapted to the model

The aggregation-based approach assumes that document structural information is additional evidence to the original content of XML elements It shows us how to combine content scores of an XML element in its own context and from its structurally related elements using the OWA operator and linguistic quantifiers

Following a similar idea, we assume that visual features are another way of obtaining auxiliary evidence to the original meanings of shots We continue to consider that the visual features of each shot can be assigned to a term vector whose value expresses the shot's additional semantic content In order to do so, we propose to employ a K-means shot clustering algorithm to organise shots of each video at the primitive visual feature level and assume that shots within the same cluster are similar visually but also semantically to some extent Terms collected from

transcripts of shots in a cluster can be linked to the member shots by their proximity to the corresponding cluster centroid

Text remains an important way of how we organise, store and convey complex ideas In video retrieval, the main advantage of full text search is its ability to represent an abstraction by both general and concrete instances at different levels of complexity But not all shots of video are accompanied by text Furthermore, even though some shots are tagged to some text, the visual perception of shots by audiences and the semantic meanings of the accompanying text may not be a match As a result of this, approaches for matching of visual features have been developed to capture users' visual information needs

Our video retrieval model for MPEG-7 will accept both text and visual queries We propose a method to utilise the outputs of the K-means shot clustering to assist the matching of visual primitives A query image can be assigned to query terms based on its proximity to cluster centroids in the collection The derived query terms can be combined with the original text query to produce the final query text Video retrieval can be carried out simply based on the dot products between the aggregated text representations of shots and the combined query

In the next chapter, we extend Kazai et al's aggregation-based approach for XML document retrieval in order to accommodate the matching of both text and primitive visual features for MPEG-7 video retrieval Our experiments of the proposed video retrieval model were carried out on the TRECVID2002 and TRECVID2003 collections and the results will be given in Chapter 6 and Chapter 7, respectively

Chapter Four

Aggregated Feature Retrieval for MPEG-7

In the previous chapter, we have shown that the aggregation-based approach for structured document retrieval is able to model the textual representations of a component in its own context and that of its structurally related components The assumption of the approach is that a document's structural information is additional evidence to the original content of XML elements We adopt this approach to model both textual and visual representations of MPEG-7 descriptions based on a similar assumption that primitive visual features are auxiliary information to the original semantics of video shots

A simple way of combining textual and visual descriptions into one single description for a shot is to construct only one vector representation by appending the visual vectors to the term weight vector There are two main problems with this method The first problem is the different similarity functions used the dot product for text features in structured document retrieval and the L1/L2 distance for visual features in MPEG-7 The second problem is the different levels of semantic representation about video content Text is a precise representation and can tell us what has just been spoken in the video content Visual features however offer us limited semantic information and are restricted to telling us what colour and texture is on the screen or if the video contains a particular range of limited concepts (i e "car", "landscape/cityscape")

Our solution for combining both textual and visual descriptions is to align a video retrieval process to a text retrieval process based on the TF*IDF vector space model via clustering of visual features The assumption is that shots within the same cluster are not only similar visually but also semantically to a certain extent, which is

detailed in section 4 1 We map the visual features of each shot onto a term weight vector by their proximity to the corresponding cluster centroid and the terms are collected from transcripts of all member shots in the cluster This vector is then combined with the original text descriptions of the shot to produce the final searchable index Sections 4 2 and 4 3 discuss the preparation of the index and query respectively A summary of the retrieval process is listed in Section 4 4

4.1 The Assumption

Query topics can range from a search for general and abstract things (i e definitions of information) to concrete and specific objects that can be seen, such as people, actions, locations or their combinations Topics for searching a video collection are particularly defined to look for concrete objects and come with textual descriptions of the information need as well as one or more video clips and/or still images to aid illustration The relevance judgements for the video collection are made based on whether the objects described in the topic can be seen on the screen The dialogue (i e ASR transcripts or closed-caption) in videos tells us what has been said but very often not what we can see on the screen A video retrieval system is encouraged to integrate primitive visual features to find "difficult" relevant shots that can not be found in the dialogue text

As for a given topic, if relevance judgements can be made for the whole video collection, we can distinguish two types of shots relevant and non-relevant If procedures are available for further separating the relevant shots, we can obtain three types of relevant shots according to the query types that are contributed to the relevance judgements, for instance, solely by visual features of image examples or by the dialogue A four-celled figure (Figure 4-1) is set up to study possible partition of the video collection based on two different query types, labelled *text query* and *visual query* Each cell interprets the query types based on which relevance judgements can be made

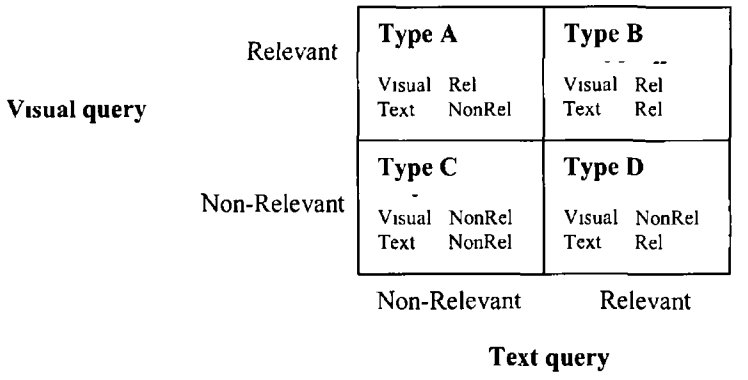


Figure 4-1 Partition of a video collection based on query types

- *Type A* – Relevance of this type can not be determined by dialogue Shots are considered to be relevant to the given topic solely based on the similarity of visual features of image examples, that is the shots contain things that can be seen on the screen but have not been mentioned in the dialogue For example, the shower scene in Alfred Hitchcock's "Psycho" Marion is murdered while showering by a mysterious women There is no dialogue to indicate the murder has taken place in the scene

- *Type B* – Relevance of this type can be decided by either dialogue or visual features It can be seen as the overlap of type A and type D Shots contain things that can be seen on the screen also has been mentioned in the dialogue Another example, such as Gene Kelly's dancing in the rain with an umbrella scene in the movie "Singin' in the Rain" where the dialogue (words in the song) indicate he is singing and dancing in the rain

- *Type C* – Shots of this type are non-relevant to the topic neither by dialogue nor by visual features

- *Type D* – Shots of this type are considered non-relevant to the topic because things satisfying the information need can not be seen on the screen even though they have been said in the dialogue Shots are useful in an interactive video retrieval system since users can examine other items that are around it An example is given in the opening scene of the movie "Citizen Kane" where Charles Foster Kane's is gasping the

famous word "rosebud" with his last breath as the camera zooms in to his mustached lips and nothing associated with the identity of the word is shown on the screen

Shots of type *B* and *D* could be returned using only text queries and those of type *A* and *B* could possibly be found using only image examples Type D shots are not normally considered relevant in video retrieval, which normally focuses on finding visual objects In fact, shots of type *A* and *B* are the ones that an automatic video retrieval system attempts to retrieve The integration of visual features into a video retrieval system is done in order to find "difficult" relevant shots such as type *A* that can not be returned from a text search of the dialogue

Shots of type *A* and *B* differ in the way that they characterise the similarity between shots and a given query The relevance of type *B* shots can be judged based on the similarity between the dialogue and a given text query The relevance of type *A* shots can only be determined by visual similarity

Type *A* and *B* shots are similar in that both are visually relevant to the image examples If we are able to group type *A* and *B* shots together via clustering of the collection using visual features, the dialogue of type *B* shots can provide auxiliary information to the existing dialogue for type *A* shots and vice versa If we aggregate the auxiliary information with the original dialogue, the "difficult" relevant shots like type *A* can be retrieved by simply calculating the similarity between the aggregated representations and a given text query

The clustering of shots can be carried out by two different methods The first one is based on the similarity of dialogue which can be implemented using text clustering and the possible outcome is to cluster shots of type *B* and *D* together The emphasis of the second method is on the use of visual features for clustering and the output is that shots in the same cluster are visually similar Shots of type *A* and *B* can be expected to group together We therefore chose the second shot clustering method that uses visual features alone to generate the clusters

Shot clustering using visual features is an unsupervised technique to group shots together that present similar visual patterns In fact, it is impossible to have shots in the same cluster that are all alike and often a cluster may include shots that are very different semantically from the majority of members Two elements can be suggested to characterise the properties of the created clusters, *visual* and *semantics* The *visual* expresses the degree of similarity of shots within a given cluster in terms of visual primitives The *semantics* concerns the degree of shot similarity in terms of text features, namely what has been said in the dialogue Again a four-celled figure (Figure 4-2) can be provided to explain four different cluster types

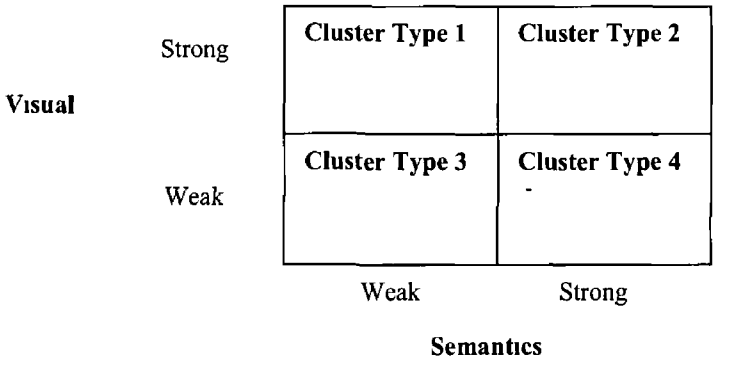


Figure 4-2 Four different cluster types

- The shot similarity of cluster type *1* is strong *visually* but offers weak *semantics* The strength of the original dialogue of each shot could *probably* be improved after the aggregation of auxiliary information provided by its members
- The shot similarity of cluster type *2* is strong both *visually* and *semantically* The strength of the original dialogue of each shot can be enhanced by its members
- The shot similarity of cluster type *3* is weak both *visually* and *semantically* The strength of the original dialogue of each shot could be reduced by its members
- The shot similarity of cluster type *4* is weak on visual but strong on semantics A cluster of type *4* is more like a text-based cluster in which patterns in dialogue are found dominant

The purpose of shot clustering based on visual features is to obtain additional information for each shot of a cluster based on the dialogues of other members in the cluster A cluster of type *2* is an ideal situation of our assumption that shots within the same cluster are similar visually but also semantically We should be aware that clusters of type *3* can be expected to occur and they are thought to be noisy clusters in that visual and semantic similarity amongst member shots is weak

4.2 Index Preparation

So far we have studied the relationship between query types and shot relevance in video retrieval The relevance judgements are made based on whether the things described in the topic can be seen on the screen Visual queries are often used to find "difficult" shots like type *A* that are not returned by any text queries Due to the visual similarity among shots of type *A* and *B*, shot clustering using visual features attempts to create clusters which contains shots from both types

An assumption is therefore given to make use of the visual and semantic similarity of shots within the same cluster The dialogue of member shots can provide auxiliary information to the original meanings of each shot Following this, we align a video retrieval process to a text retrieval process based on the TF*IDF vector space model via clustering of visual features [Ye & Smeaton, 2003] Visual features of each shot are mapped onto a term weight vector based on the term occurrences in the associated cluster This vector is then combined with the original term weight vector of the shot's dialogue to produce the final searchable index We summarise our index preparation for visual features into following three steps and the details of each preparation step will be given in the following subsections

- Apply K-means shot clustering to obtain clusters for each video rather than all video in the entire collection Features considered here include colour histogram, dominant colour and edge histogram as described in MPEG-7
- Assign meanings to each cluster using a modified TF*IDF algorithm in which the indexed unit is replaced by a cluster

- Use a simplified Bayesian approach to derive the text description of each shot based on its cluster meanings

Having obtained a text description for each shot via clustering, we aggregate this with the original term weight vector for each shot to create its final term weight vector using the Ordered Weighted Averaging (OWA) operators and linguistic quantifiers

4.2.1 K-means Shot Clustering

Clustering is an "unsupervised classification" in which class labelling of the training data items is not available The process puts data items into subsets only by their actual observations items in the same cluster are alike and items from two different clusters are dissimilar To be more specific, if each of the data items in a collection have a vector representation with a length of l and can be viewed as a point in the l-dimensional space, clusters are defined as "continuous regions of this space containing a relatively high density of points, separated from other such regions by regions containing a relatively low density of points" [Everitt, 1980, pp60]

Hierarchical clustering and partitional clustering are two main types of clustering methods A hierarchical method is a sequence of nested groupings created from a proximity matrix [Anderberg, 1973] The method works in a bottom-up way by first putting each data item in its own cluster and constructing a dissimilarity (or similarity) matrix for all distinct pairs of data items It then repeatedly merges a pair of clusters that has the minimum distance into a single cluster by a graph edge and updates the dissimilarity matrix until all items are connected in the graph The dissimilarity matrix is updated by removing the rows and columns of the pairwise clusters and adding a row and a column for the newly merged cluster The output of the method is a dendrogram that can be broken horizontally at different dissimilarity levels to generate different groupings of the data

The hierarchical method is versatile and it works well on data sets that are well-separated, chain-like or have concentric clusters [Jam et al, 1999] It is efficient with small data sets since the use of a proximity matrix in the method limits the number of data items that may be clustered [Anderberg, 1973]

A partitional clustering method generates a single partition of the data items via a pattern matrix The basic idea begins with some initial partition of the data items and repeatedly modifies cluster memberships by moving items from one cluster to another in order to improve a criterion function, until a global or a local optimum is found [Jain & Dubes, 1988] In a global optimum, each cluster is represented by a prototype and the items are assigned to clusters based on most similar prototype, for instance, clusters can be identified as high-density regions in the feature space A local optimum is reached by local structure in the data items, for example, a cluster can be constructed by assigning an item and its k nearest neighbours to the same cluster

The most frequently used criterion function in partitional clustering methods is the square-error clustering criterion The intuition is that a cluster is a set of items whose inter-pattern distances are smaller than inter-pattern distances for items not in the same cluster The cluster can be visualised as a region of high density (or a hypersphere) in the feature space

The square-error clustering criterion can be formulated as follows given n data items in an l-dimensional pattern matrix, we determine a partition of the items into K clusters $\{C_1, C_2, \ldots C_K\}$ such that the items in a cluster are more similar to each other than to items in different clusters Each item is assigned to only one cluster and no cluster can be empty The centroid of cluster C_j is the mean vector of the members in the cluster,

$$m_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij}$$
$$\sum_{j=1}^{K} n_j = n \qquad (4\text{-}1)$$

where x_{ij} is the i-th data item belonging to cluster C_j and n_j is the number of items in cluster C_j The square-error for cluster C_j, also known as the within-cluster variation, is the sum of the squared Euclidean distances between items in C_j and its cluster centre m_j

$$dist_i = (x_{ij} - m_j)^T (x_{ij} - m_j) \quad (4\text{-}2)$$

$$e_j^2 = \sum_{i=1}^{n_j} dist_j \quad (4\text{-}3)$$

The square-error for the all K clusters is the sum of the within-cluster variations

$$E_K^2 = \sum_{j=1}^{K} e_j^2 \quad (4\text{-}4)$$

The problem of a square-error clustering method is considered as the problem of minimising E_K^2, for fixed K The resulting K clusters are made as compact and separated as possible The most popular of the square-error clustering methods is the K-means method due to its straightforward implementation [Anderberg, 1973] [Jain & Dubes, 1988] The K-means clustering algorithm can be summarised as follows

1 Randomly choose K data items as an initial clustering of the n items and the feature vectors of the K chosen items are initialised to be the cluster centroids

2 Assign each data item to the closest cluster centroid with Formula 4-2

3 Recalculate the cluster centroids using the current cluster memberships with Formula 4-1

4 If a convergence criterion is not met, go to step 2 The convergence criterion is often defined as the minimal decrease in squared-error (Formula 4-4) For easy computation, a popular convergence criterion used is no re-assignment of items to new cluster centroids

Some of the parameters in the K-means clustering method are reviewed and taken into account in our work

- *The number of clusters K*

This needs to be explicitly specified by users In order to find the best partition, the cluster validity analysis can be employed to find a clustering that gives minimum intra-cluster distance while keeping the maximum inter-cluster distance [Jain & Dubes, 1988] The cluster separation measure for K clusters is defined as $\lambda(K)$

$$\lambda(K) = \frac{1}{K}\sum_{i=1}^{K} \max_{1 \le n \le K} \{\frac{e_i^{\,2} + e_j^{\,2}}{\beta_{ij}}\} \tag{4-5}$$

where

$$e_j^2 = \sum_{i=1}^{n_j} (x_{ij} - m_j)^T (x_{ij} - m_j) \tag{4-6}$$

$$\beta_{ij} = (m_i - m_j)^T (m_i - m_j) \tag{4-7}$$

e_j^2 is the intra-cluster distance of cluster C_j and β_{ij} is the inter-cluster distance between cluster C_j and C_i If we run the clustering algorithm for different K values from a set of integer value $\{K\}$, the best K can be selected as the one that has the minimum cluster separation measure [Ngo et al, 2001]

$$K_{opt} = \min_{k \in \{K\}} \{\lambda(k)\} \tag{4-8}$$

- *Initial partition*

The outcome of the method is sensitive to the selection of the initial partition A local minimum of the criterion function can be obtained if the initial partition is not properly chosen To overcome this, the algorithm should be run with different initial partitions to obtain the best clustering [Jain & Dubes, 1988] If they all result in the same final partition, a global minimum

of square-error could be reached The process can be considered after we obtain the best number of clusters, K

- *Convergence*

The K-means clustering algorithm does not guarantee that a global minimum can be achieved A maximum number of iterations is required to be specified to prevent infinite loops We specified the maximum number of iterations to be 300

The K-means clustering method is convergent because the successive created partitions present a strictly decreasing sequence of square-error values E_K^2 (Formula 4-4) [Anderberg, 1973] A data item is re-assigned only if it is closer to the new centroid of the gaining cluster than the old centroid of the losing cluster If Euclidean distance is the chosen distance measure, the within-cluster variation e_K^2 (Formula 4-2) about the old centroid of the losing cluster decreases more than it increases for the gaining cluster Hence, the overall sum of square-error E_K^2 about the new centroids for the partition decreases Each successive partition has lower value E_K^2 than their direct predecessors do

Due to the large number of data items and the size of feature vectors used for each data item in our experiments, we chose the L1 or "Manhattan" distance function as the measure of convergence in the K-means clustering in order to speed up the process The overall within-cluster variation is defined as [Anderberg, 1973]

$$E_K = \sum_{j=1}^{K} \sum_{i=1}^{n_j} | x_{ij} - m_j | \qquad (4\text{-}9)$$

where m_j is the centroid vector of cluster C_j, in which each cell m_{jl} is the median value of the l-th feature variable for cluster C_j The clustering method is revised to compute the cluster median and minimises the overall within-cluster absolute errors This method can be shown to be convergent using the same argument as above

Following this, we employ the K-means clustering method to video collections to group shots with similar colour and edge features for each single video programme Three visual features defined in MPEG-7 (see chapter 2 3) are under consideration for each shot (1) 9 region * dominant colour in the RGB colour space, (2) 4 region * 16 bin colour histogram, (3) global 80 edge component histogram All three visual features are normalised into a range between 0 to 1

Colour has apparently been the most prevalent of low-level visual features in image and video IR and dominant colour, combined with the 9-region colour histogram, seems to give a balance between allowing a dominant colour in a frame (such as green on a football pitch or red or orange in a sunset) and incorporating colour distributions throughout the (four) regions of the frame Edges, and the presence of an edge histogram is a useful measure to capture the number of "things" or lines in an image, where a city scene with cars and buildings and roads and people etc will have many lines between objects, hence edges, whereas a shot of a single vase against a plain background will have few lines

Given any two shots of each single video programme within the video collection, noted as $Shot_i$ and $Shot_j$, each shot has its corresponding visual feature vectors $f^{(m)}$ (M is the number of features and $1 \leq m \leq M$) and each feature vector has a length of l Based on the Manhattan distance, the distance $Dist$ between $Shot_i$ and $Shot_j$ is simply a linear combination of the different visual features, as defined in Formula 4-10 The absolute error in Formula 4-9 can be substituted by value $Dist$

$$Dist(Shot_i, Shot_j) = \sum_{m=1}^{M} \sum_{p=1}^{l} | f_{pi}^{(m)} - f_{pj}^{(m)} | \qquad (4\text{-}10)$$

We apply a sliding window technique in the partition update process for the K-means shot clustering algorithm because of the temporal characteristics of videos If, for any of the candidates, the adjacent shots of a given shot $Shot_i$, are similar enough to the shot, it should be put into the same cluster as $Shot_i$ Given a window size WIN of 2, for instance, the next 2 shots of a shot in the sequence would be the candidates A threshold T is required to determine the "similar enough" criteria But there is no immediate way of setting the threshold without the knowledge of

distribution of the dissimilarity values between shots An indirect method to assign the threshold is described for each video as below

1) Given a window size *WIN*, calculate the dissimilarity values between a shot and its candidate shots If a given video contains *N* shots, there would be *WIN* * *N* dissimilarity values in total

2) Sort the *WIN* * *N* dissimilarity values in ascending order Shots are similar enough if their dissimilarity values are within the range set by the top *xT* percent of all sorted dissimilarity values The dissimilarity value in position *WIN*N*xT* is the threshold *T* for the video

When the window size *WIN* is set to 0, our shot clustering procedure is equivalent to a traditional k-means clustering process Having obtained the similarity threshold *T* when *WIN* $\geq$ 1, our shot clustering is carried out in the following steps

1 Randomly choose *K* shots as an initial clustering of a given video programme and the feature vectors of the *K* chosen items are initialised to be the cluster centroids C_k ($1 \leq k \leq K$)

2 Assign each shot $Shot_i$ of the video programme to the closest cluster centroid C_k with Formula 4-10 If, for any of the candidates, the next *WIN* shots of $Shot_i$ in the sequence, are similar enough to $Shot_i$, it should also be put into cluster C_k

3 Recalculate the cluster centroids C_k ($1 \leq k \leq K$) using the current cluster memberships with Formula 4-1

4 If a convergence criterion is not met, go to step 2 For easy computation, we used a popular convergence criterion, namely no re-assignment of items to new cluster centroids

The results of a cluster partition do not suggest a "best" structure for all data items because no precise definition of "cluster" exists More often one or more clusters lead to structures (observations) on only part of the data As a result of this,

certain clusters may be well formed once uncovered so that the structure of the data are likely to be revealed, meanwhile, totally unexpected aspects of structures might be discovered in the process (see section 4 1 on the different cluster types)

4.2.2 Assigning Meanings to Clusters

Having obtained a cluster partition based on the visual similarity, we can assign a semantic representation (i e a term weight vector) to each cluster based on the premise that the similarity of shots in the same cluster The semantic representation consists of (1) terms collected from the automatic speech recognition (ASR) of words spoken in the member shots for the cluster and (2) weights calculated based on a modified TF*IDF weighting schema The spoken terms are the only source of semantic information for our experiments although other sources are possible such as OCR of any printed captions or text appearing in the video [Sato et al, 1998]

One of the earliest text information retrieval systems based on a vector space model was the SMART system at Cornell University [Salton, 1983] The semantics of a document is represented as a one-dimensional vector, where the position of each cell corresponds to a processing term token and the value of the cell is weighted based on the frequency of occurrence of the token in the document A text query can thus be represented as a vector in the same dimensional space The dot products between the query vector and document vectors provide a basis for determining the ranking of the relevant documents

One popular weighting schema for the vector space model is the TF*IDF schema based on the observation that the frequency of occurrence of distinct word forms (i e tokens) in a document denotes the importance of these words for content representation [Luhn, 1958] The variants such as *organise*, *organisation* and *organiser* are often regarded as the same word containing identical meanings If a word occurs in every document in the collection, its value for distinguishing one document from another will be almost nil whereas if a word occurs in very few documents in the collection its value for distinguishing will be much greater Therefore, the weight W_{ij} of a given token T_j in a document D_i is proportional to TF_{ij}

(i e the frequency of its occurrence in the document) and inversely proportional to IDF_j (i e the frequency of its occurrence in the whole collection)

$$W_{ij} = TF_{ij} * [Log_2(n) - Log_2(IDF_j) + 1] \quad (4\text{-}11)$$

where n is the number of documents in the collection

We slightly modify the TF*IDF weighting scheme to index our cluster partition by assigning word tokens to clusters A cluster rather than a document is treated as the indexing unit TC_{kj} is the frequency of occurrence of a token T_j in a given cluster C_k CF_j gives the number of unique shots in the collection that contain token T_j nc indicates the total number of clusters created in the search collection The weight of token T_j in cluster C_k is defined in Formula 4-12

$$W_{kj}^{(cluster)} = TC_{kj} * [Log_2(nc) - Log_2(CF_j) + 1] \quad (4\text{-}12)$$

If a cluster C_k can be summarised by a term weight vector W_k = {W_{1k}, W_{2k}, , W_{mk}}, where m is the total number of term tokens in the collection, a term-by-cluster matrix can be formed to record the degree of associations between any clusters and term tokens in the collection (see Figure 4-3) The following comparison results can be obtained by observing the matrix

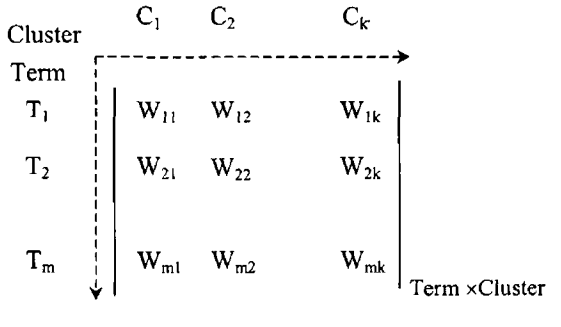


Figure 4-3 A term-by-cluster matrix

- *Term-cluster* comparisons are the values for the individual cells W_{ij} of the matrix It estimates the extent to which a term T_i and a cluster C_j are associated with each other

- *Cluster-cluster* comparison is the dot product of two vector columns of the matrix It estimates the extent to which two given clusters have a similar set of terms
- *Term-term* comparison is the dot product of two vector rows of the matrix It estimates the extent to which two given terms have similar patterns of occurrence across the cluster set

A problem with the process of assigning meanings to clusters is the sensitivity to the cluster partition As described in section 4 1 if a cluster like type *A* or *B* is obtained which has shots that are similar visually and semantically, the values of assigning meaning to the cluster may be useful On the contrary, if a cluster like type *C* is returned in which shots show no similarity pattern at all, the meaning of the cluster is useless We should also bear in mind that as the number of shots within a cluster increases the number of terms used to describe the cluster also increases and as a consequence the major meanings of the cluster might become less prominent

4.2.3 Deriving Text Descriptions for Shots Based on Cluster Meanings

The purpose of the process is to enhance the meaning of a shot by adding a certain level of semantic information based on the cluster meanings The original ASR transcripts associated with a shot are often limited We expect that the introduction of cluster meanings would help us understand more about the shot based on the assumption that shots within the same cluster present weak semantic similarity

The temporal characteristic of videos allows us to add meaning to a shot based on the shots that are around it Such a way of enhancing the meaning of a shot is much easier to implement than our proposed method in which a clustering algorithm is required What our method attempts to solve is to include the meanings (i e terms) from shots that are similar within a video to a shot These shots are not necessarily neighbours to the shot They could be from either the beginning of the video, the middle or the end

Another advantage of this approach is that silent shots from videos which have no ASR could receive some semantic terms from their similarity to shots in the same cluster which have an ASR associated with them There are times when shots within some groupings are not semantically similar Naturally shots in these groupings will not benefit from the addition of cluster terms, but would create misleading information about the shot

The K-mean shot clustering method is a hard-clustering in which the shot-cluster memberships take up two values {0, 1} We can not estimate the degree to which a given shot falls into a cluster Instead only the distance values between the cluster centroid and the member shots are available We attempt to approximate the weight of a term for a given shot by considering the corresponding cluster-shot distance If shot i is closer to the cluster centroid than shot j, the weight of a term for shot i is higher than shot j Let $Dist_{ik}$ be the distance between a given shot S_i and its corresponding cluster centroid C_k The inferred weight $SW_{ij}^{derived}$ of term W_j for shot S_i is defined in Formula 4-13

$$SW_{ij}^{derived} = (1 - Dist_{ik}) * W_{jk} \qquad (4\text{-}13)$$

where W_{jk} is the weight of term W_j in cluster C_k

4 2.4 Aggregation of Shot Text Descriptions

In order to combine text annotations and visual features into one single MPEG-7 description, we have summarised the visual features of each shot as a term weight vector based on the similarity pattern discovered in its cluster Two types of text information for each shot are ready for aggregation (1) the original ASR transcripts obtained directly from the text annotation in the MPEG-7 description, (2) the derived text information from the visual features, also regarded as the auxiliary information to the shot

The OWA operator and "*at least ½*" linguistic quantifier is chosen as the aggregation method in this work (see Formulas 3-1 and 3-2) and we use *Agg()* to denote the method in the rest of the thesis The process combines the original and derived term weights of a shot according to their associated importance values which is specified by a linguistic quantifier The chosen quantifier "*at least ½*", defined by $Q(r) = r^{1/2}$, emphasises high weights by placing more importance values to them

The aggregated weight of term T_j of a given S_j is given as SW_{ij}^{agg}

$$SW_{ij}^{agg} = Agg(SW_{ij}, SW_{ij}^{derived}) \qquad (4\text{-}14)$$

where SW_{ij} is the original term weight and $SW_{ij}^{derived}$ is the derived term weight from the meaning of the cluster which contains S_j

Given a term for a shot, we assume the original and derived term weights are $W_1 = 6$, and $W_2 = 8$ The sum of term weights is 14 ($T = W_1 + W_2$) The ordering of the term weights and the associated importance value α_j are

	b_j	α_j
W_2	8	$Q(^8/_{14}) - Q(^0/_{14}) = 0\ 756$
W_1	6	$Q(^{14}/_{14}) - Q(^8/_{14}) = 0\ 244$

The aggregated term weight can be obtained as follows

$$W = Agg(W_1, W_2) = \sum_{j=1}^{2} \alpha_j b_j = (\ 756)(8) + (\ 244)(6) = 7\ 512$$

The aggregated weight W has a bound between W_1 and W_2 If the original term weight W_1 is lower than the derived weight W_2, the aggregated weight W will have a higher value than the original weight W_1 and adding extra meaning to the shot via clustering is considered useful

On the contrary, if the original term weight were higher than the derived weight, the aggregated weight would decrease For example, W_1 = 8 and W_2 = 6, the aggregated weight W decreases to 7 512 In such cases, additional meaning from the cluster partition to the shot is not useful since the final weight is weakened Therefore, we use quantifier "*at least ½*" to emphasise high weights and allow for a marginally decrease in the aggregated weight in comparison to the original weight

If the original term weight W_1 has 0 value and the derived weight W_2 does not equal to 0, the aggregated weight W is given the same value as W_2, indicating that the extra meaning is added into the shot But no attention is paid to whether the extra meaning can enhance or weaken the original shot meaning

4.3 Query Preparation

Having aggregated text annotations and visual features into one single MPEG-7 description, we are ready to study how to carry out queries for video retrieval via the aggregated description Two types of video queries are under consideration (1) text-only, (2) non-text (i e consisting of an image/ video clip represented by a key frame) Low-level visual features are calculated for a non-text query including (1) 9 region * dominant colour in the RGB colour space, (2) 4 region * 16 bin colour histogram, (3) global 80 edge component histogram

The conventional way of handling a non-text query in video retrieval is to compare and rank the distances between a given non-text query and all searchable units in the collection The searchable unit can be a shot or a cluster centre if shot clustering is applied In our work, the searchable unit is the cluster centre Given a non-text query, if the *topK* closest clusters are returned (where *topK* is the number of the closest clusters to the query), the shots within the *topK* clusters would be very similar to the query in terms of visual features

Based on our assumption that shots in the same cluster have a certain degree of visual and semantic similarity, the meanings gathered from the *topK* clusters can be regarded as a text description of the given non-text query The text description can

be used in the same way as a text-only query to return relevant shots via the aggregated index using the vector space model

Generating a text description of a non-text query is similar to an automatic query expansion process in text retrieval The traditional view of automatic query expansion is to expand a given text query based on the terms found in the top k relevant documents returned The use of previously retrieved document texts suggests additional or alternative possible query terms Following the similar idea, we use a non-text query for query expansion in order to find additional query terms based on the texts of the *topK* similar clusters to the query The *topK* clusters are withdrawn after its text description is obtained and the text description continues the search Listed below are the steps showing how to map a non-text query onto a text query

(1) Find the *topK* most similar clusters C_m ($1 \leq m \leq K$) to a non-text query $Q^{non\ text}$ based on visual features *topK* is the number of the most similar clusters The similarity between a cluster and non-text query is determined by the distance between the cluster centroid and the query, noted as $D_m(Q^{non\ text}, C_m)$ (see Formula 4-10)

(2) Take the term vector of the *topK* chosen clusters and add them together to form a single query term vector based on the normalised inverse distance NID_m The normalised inverse distance is inversely related to the distance D_m because the closer a selected cluster centroid is to the non-text query, the more importance value is given to the query terms collected from the cluster The weight of query term T_j, noted as $QW_j^{non\ text}$, is given in Formula 4-15, where $W_m^{(cluster)}$ is the weight of token T_j in a given cluster C_m

$$QW_j^{non-text} = \sum_m W_{mj}^{(cluster)} * NID_m(Q^{non-text}, C_m) \qquad (4\text{-}15)$$

$$NID_m = 1 - \frac{D_m}{\sum_{m=1}^{K} D_m} \qquad (4\text{-}16)$$

When more than one non-text query examples are used in a query, for instance, multiple image examples including the key frames of video examples, each non-text query $Q_a^{non\ text}$ ($1 \le a \le S$, where S is the number of image examples) is first mapped to a query term vector $QW^{non\ text}{}_a$ based on visual features as described in Formula (4-15) We then combine these a query vectors to form a single query term vector for all the non-text queries using the aggregation method as shown in section 4 2 4 and Formulas 3-1 and 3-2

$$QW^{content} = Agg(QW^{non-text}{}_1, \ , QW^{non-text}{}_S) \tag{4-17}$$

Given two different types of queries for a given topic (i e text-only, non-text query), an attempt was made to join them together into a single query form, namely a text query We aggregate two query term vectors (1) the original text-only query $QW^{text\ only}$, (2) a derived text query $QW^{non\ text}$ from the non-text queries The aggregated weight of query term T_j is given as QW_j^{agg}

$$QW_j^{agg} = Agg(QW_j^{text-only}, QW_j^{non-text}) \tag{4-18}$$

An outcome of using a derived text description of non-text queries is the possibility of bringing more unrelated terms to a given topic It thus loosens the specificity of the topic degrading the system performance Original terms in text-only queries play an important role in video retrieval Prior to aggregating two different types of queries, it is suitable to assign more important values to the original query terms than the derived terms But how much more weight we should be assigned to the original query terms is not clear and perhaps might not always be consistent

The variable PF is introduced in order to maintain the extent of topic specificity but also to include additional information from the derived terms The degree of impact of original text-only queries on the retrieval performance can be estimated when including non-text queries The original query term weight $QW_j^{text\ only}$ is given in Formula 4-19, where W is a binary value indicating the presence or absence of a

term The *max()* function finds the maximum value of the derived term weights $QW_j^{non\ text}$

$$QW_j^{text-only} = W + \max_j (QW_j^{non-text}) * PF \quad (4\text{-}19)$$

4.4 Retrieval

Retrieval can be done in a straightforward manner by calculating the dot products between the aggregated query term vector and the aggregated index vector for each shot and sorting the dot products in descending order The dot product between an aggregated query and shot S_i is given in Formula 4-20

$$Sim_i = \sum_j QW_j^{agg} * SW_{ij}^{agg} \quad (4\text{-}20)$$

where j is the order position of the j-th index term

A typical video retrieval approach is to search different MPEG-7 descriptions separetely and to combine ranked results from the different searches using a sum weighted method Our approach attempts to integrate the different descriptions into the index and query preparation stages - no combination of ranked results is required Figure 4-4 below shows the data flow diagram of our video retrieval system for MPEG-7

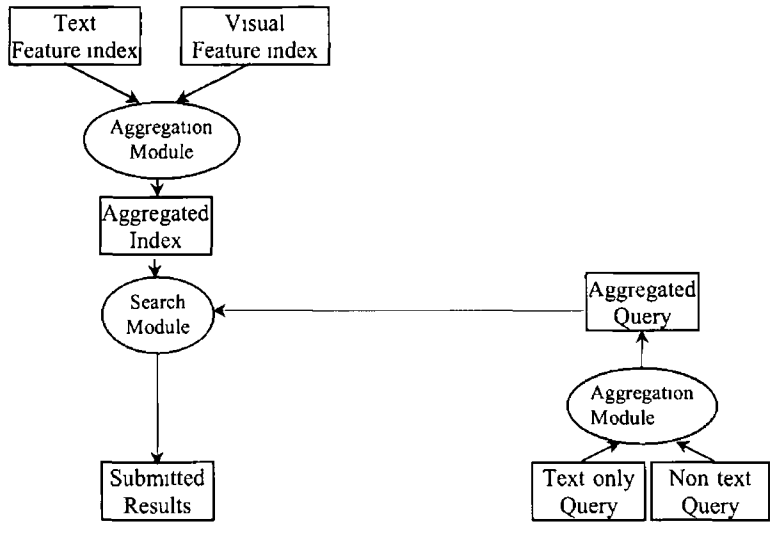


Figure 4-4 Data flow diagram of our video retrieval system

4.5 Conclusions

The design of this video retrieval approach is centred on the notion of the hierarchical structure (key frame, shot, scene/story and video) and two related, complementary types of information (text and visual features) that complete the definition of what we consider a good basis for video representation Our study of the hierarchical structure is limited in the use of a key frame at a shot level but a more elaborate shot-shot matching would have to account for the temporal aspect and cater for object movement and camera movement and needs more than colour histograms and edges as we use here This kind of simple structure that we use to compute shot-shot matching is intensively used in many applications, in particular those defining a shot as the default retrieval unit

Much effort in this work was put on the integration of the two complementary features of videos rather than on extending either, features we consider required to deal in different retrieval scenarios Full text search can be performed to retrieve shots whose accompanying dialogue (e g ASR transcripts) is relevant to the text However the returned shots may not all be relevant as the dialogue tells us what has been said but not necessarily what we see on the screen Image-based search using visual features can be carried out to locate "difficult" relevant shots that would not be found by a dialogue search

In this chapter, we described an approach to index MPEG-7 descriptions for video retrieval extended from the aggregation-based method for structured document retrieval The assumption of the aggregation approach is that a document's structural information is extra evidence of content that can be used during the indexing process We modified the idea slightly for video retrieval whereby the visual features of shots are seen as the auxiliary evidence that enhance the shots' original meanings

The main thrust of our approach focuses on mapping a video retrieval process to a text retrieval process based on the TF*IDF vector space model via clustering of low-level visual features We assume the visual and semantic similarity of shots within the same cluster The semantic similarity is considered to be weak because the cluster does not result directly from a text clustering process, but indirectly from clustering of visual features We should be aware that it is impossible to have shots in the same cluster that are all alike and more often a cluster may include shots that are very different from the majority of members

Having obtained a cluster partition for each video in the collection, we can map the MPEG-7 visual descriptions of each shot onto a text description thereby providing auxiliary information to the original MPEG-7 textual description of the shot The text description is treated as a term weight vector which can then be combined with the original MPEG-7 textual description to produce a final searchable index Shots in well-formed clusters will benefit from the addition of clustering meanings while poorly-formed clusters would create totally unexpected information about the member shots The advantage of the method is the inclusion of semantic meaning (i e terms) from shots that are similar visually within a video to a shot These shots are not necessarily neighbours to the shot They could be from either the beginning, middle or end of a video

Our video retrieval method introduces a way of using text-only and non-text queries together for retrieval via an aggregated index We map a non-text query onto a text query description and combine it with the original text-only query for video retrieval The mapping process is problematic since it allows a non-text query to

expand into a good description closely related to the topic or into a poor description completely different from the original topic

Experiments were carried out on the TRECVID2002 and TRECVID2003 search test collection in order to evaluate the assumption both in the index and query preparation stages, a description of which will be given in Chapters 6 and 7 respectively Query optimisation for locating relevant shots with least estimated time and cost is not covered in this work No query language is taken into account since we are more concerned with searching the MPEG-7 content, in particular at the shot level, than querying the MPEG-7 structure

Chapter Five

Video Retrieval System Evaluation – TREC Video Track

The previous chapter of this thesis discussed the requirements and design of a video retrieval system and in this chapter we will look at evaluation of retrieval efficiency and effectiveness Retrieval *efficiency* emphases the minimum effort, time and cost for users and computers The computer cost can be easily estimated, but the user effort is a little more difficult to quantify Retrieval *effectiveness* is concerned with the ability of a retrieval system to find a large number of relevant documents while reducing the number of irrelevant documents returned Retrieval effectiveness is often measured in terms of recall and precision of search outcomes The computation of recall and precision is covered in detail in section 5 2

Up to 2001 the evaluation of video retrieval systems was carried out mostly by academics using a few small, well-known video collection corpora, their own manually annotated content [Browne et al, 2000] or even smaller test sets such as the MPEG-7 video content set [MPEG7-W2466] In 2001 the annual Text Retrieval Conference (TREC) sponsored by the Defence Advanced Research Project Agency (DARPA) and the National Institute of Standards and Technology (NIST) created an annual workshop (TRECVID) dedicated to the research into digital video retrieval The TRECVID workshop provides a standard video collection, search topics and relevant judgements for testing of system performance Section 5 3 gives more insight into the TRECVID tests

5.1 Introduction to Information System Evaluation

An information retrieval system should provide users with straightforward and efficient access to information, that is, to help them locate useful information accurately, quickly and with the minimal effort from users A system that fails to find what users want (when it is available) is not desirable, nor is a system that is difficult to use, expensive to run, or too slow to return results [Marchionini, 1995] Two main reasons for evaluating retrieval systems can be summarised as follows [Salton & McGill, 1983] [Rijsbergen, 1979]

- From a commercial perspective, the evaluation of a prototype system provides enough information so that developers can make a decision as to (1) whether they want to build such a system and (2) whether it will present significant value to the operational environment

- From an academic perspective, to determine the effects of changes made to an existing retrieval system – one might want to know (1) whether system performance improves or degrades when a particular algorithm is replaced or changed and (2) to what extent will it affect the performance

Objective measurements of retrieval efficiency and effectiveness are desirable as they are easy to obtain and usually free of personal opinions Cleverdon et al listed six critical measurable quantities [Cleverdon et al, 1966]

1 The response *time* of the system between the request being made and the judgement being given The response time is readily estimated
2 The form of *presentation* of the retrieved items It is easy to express
3 The *effort* for users in formulating a query, handling a search and making judgement on the output The user effort can be assessed partially as the time required for query formulation and interaction with the system
4 The collection *coverage*, that is, the degree to which the system includes all relevant items to a request It may not be easy to estimate if the number of items of concern in a given topic is unknown One solution is

to determine roughly the total number of items by looking up the index terms [Salton & McGill, 1983]

5 The *recall* which estimates the ability of the system to return a large amount of relevant items in answer to a search topic

6 The *precision* which measure the ability to show only relevant items amongst all retrieved items

The calculation of the recall and precision are difficult because both quantities require the determination of *relevance* for items (documents) Before introducing the computation of recall and precision in section 5 2, we now present a brief summary about the concept of relevance

Relevance is a measure of the correspondence existing between an information requirement and a document by a user [Saracevic, 1975] The notion of relevance is subjective, depending on the state of knowledge of the user at the time of the search For instance, a document may be relevant if it deals with the appropriate matter but not be relevant if the user has already viewed the content A thorough review of the factors involved in relevance is given in [Saracevic, 1996]

Relevance also takes on an objective view, making it possible to measure retrieval effectiveness This view only concerns topic relatedness rather than the subjective aspects of relevance A document is relevant to a request if it contains the minimal content that is relevant to the request [Rijsbergen, 1979] Logical relevance can be considered by the degree to which a document covers the content that is appropriate to the request It is now realised in information retrieval as a measurement function of the similarities between the request and document contents [Salton & McGill, 1983]

5.2 Recall and Precision

Precision is a measure of the accuracy of a search and is defined as the ratio between the number of relevant items retrieved and the total number of items retrieved (Formula 5-1) Precision is inversely related to the total number of retrieved items

precision decreases as the number of retrieved items increases A topic will have a maximal precision of 1 0 when only relevant items are retrieved A topic with fewer than five relevant items will have a precision less than 1 0 after five items are returned regardless of how the items are ranked

$$precision = \frac{number\_of\_relevant\_items\_retrieved}{total\_items\_retrieved} \quad (5\text{-}1)$$

Recall is a measure of the retrieval system's ability to locate all the relevant items in a collection and is defined as the ratio between the number of relevant items retrieved and the total number of relevant items in the collection (Formula 5-2) A topic will have a maximal recall of 1 0 when all relevant items are retrieved A topic with more than five relevant items will have a recall less than 1 0 after retrieval of five items

The value of the denominator (*total_relevant_items_in_collection*) is not usually available for large collection If a search collection is small enough it is possible to make reasonable accurate relevance judgements for all items in the collection with respect to each query Making relevant judgements for a large collection is time consuming An approach to reduce this work-load is known as "pooling", used in the TREC experiments (see section 5 3 1)

$$recall = \frac{number\_of\_relevant\_items\_retrieved}{total\_relevant\_items\_in\_collection} \quad (5\text{-}2)$$

Recall is not affected by total number of retrieved items based on the factors in Formula 5-2 Recall could be the same as long as five relevant items are returned regardless of the total number of retrieved items, say 100 or 1000

Precision and recall tend to be inversely related recall will usually increase as the number of retrieved items increases, meanwhile, precision is inclined to decrease A set of recall and precision value pairs can be obtained if conditioned upon the number of retrieved items For instance, one recall-precision pair can be calculated after retrieval of K items, the next pair obtained after retrieval of K+1

items, followed by that after retrieval of K+2 items, and so on up to the defined number of retrieved items, say 100

Given a set of recall-precision value pairs, a recall-precision curve can be created by plotting the precision against the recall The curve is a collection of points in a 2-dimension graph whose X co-ordinates are the recall levels and Y co-ordinates are the precision values Precision is measured by interpolation at different recall levels 0, 0 1, 0 2, , 1 are the frequently used recall levels For the calculation of a precision-recall curve, the reader is referred to [Salton & McGill, 1983]

For each query submitted to a retrieval system one recall-precision curve can be calculated, illustrating the performance of the system for the query The overall performance of a retrieval system is usually measured for a number of different topics and a recall-precision curve can also be given in which the precision is the average precision of all topics at each recall level

A precision-recall curve plots precision as a function of recall It consists of a set of ordered value pairs If i denotes a recall level (i = {0, 0 1, 0 2, 1}), P_i denotes precision at i recall level and R_i recall Pairs (P_i, R_i) can be obtained for two retrieval systems A and B

If $P_i^{A} \leq P_i^{B}$ for any i, the search results of system B is considered to be superior to those of system A Figure 5-1 shows the curves for system A and B The curve closest to the upper right-hand corner of the figure is system B and this indicates better performance

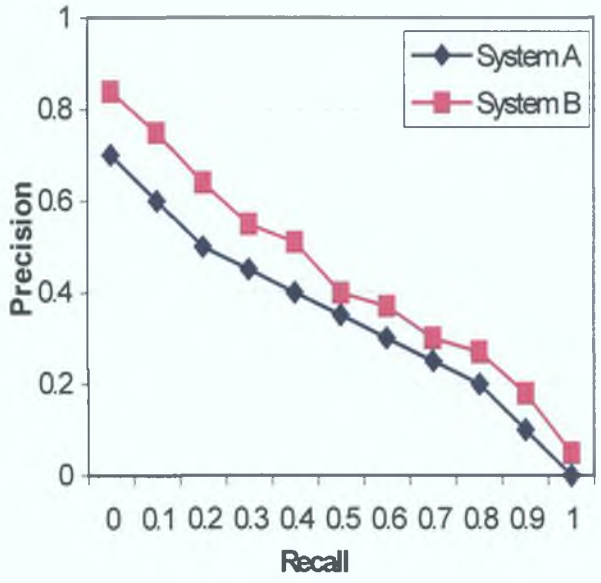


Figure 5-1: Precision at recalls for systems A and B: B is superior to A

If $P_i^{A} > P_i^{B}$ when $i < 0.4$ and $P_i^{A} < P_i^{B}$ when $i > 0.4$, or vice versa, it is difficult to determine which system performance is better (see Figure 5-2).

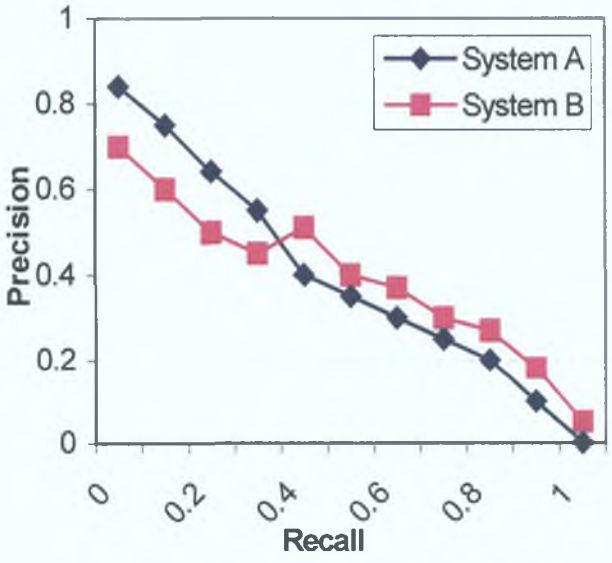


Figure 5-2: Precision at recalls for system A and B: difficult to determine which system performance is better

5.3 TREC Video Track - TRECVID

The Text Retrieval Conferences (TREC) was created by the Defence Advance Research Projects Agency (DARPA) and the National Institute of Standards and Technology (NIST) in 1992 Its goal is to provide common bases for retrieval system (initially text based) experimentation and comparison They accomplished this by providing standardised evaluation, test collections and query topics A new track dedicated to digital video retrieval, named as TREC Video Track (TRECVID), was introduced in 2001 A great deal of experience has been gained on TRECVID since the first running in 2001, as the details and focus of the video retrieval experiments have evolved

TRECVID2003 provided a set of training and evaluation data, each containing approximately 60 hours of ABC and CNN news programmes from late January through June 1998 Four types of tasks were examined (1) story segmentation, (2) shot boundary determination, (3) high-level feature extraction, (4) search

The search task for TRECVID2003 defined 25 topics each of which came with a text description to express the information need, and some image/video examples to illustrate the search topic The research groups participating in TRECVID are allowed to submit at most 1000 shots in ranked order that meet each of the search requests in one run The submissions are examined manually for relevance using the pooling method described in section 5 3 1 and recall and precision values are obtained for each run

Apart from precision and recall, other measures used in TRECVID will be given in section 5 3 2, including precision at different document cut-off levels, average precision by topic and mean average precision

5 3.1 The Relevance Judgements

It is impossible to examine the relevance of each shot with respect to each topic for a large collection, either text documents of video TRECVID adopts a pooling method to determine the possible relevant shots for each topic [Harman, 1992] The relevance judgements for a given topic are determined as follows

1 Collect the top *N* ranked shots from different submitted runs for the given topic *N* could be 50, 100 or 200 *N* is the number of items taken from each run, known as pool depth
2 Put the *N* shots from all runs into the given topic pool If there are *X* runs submitted by different researchers, there would be *X*N* shots after the combination
3 Remove the duplicate shots keeping only one copy in the pool Many shots are retrieved in the top *N* for more than one run and the size of pool is in fact smaller than *X*N*
4 An information specialist is asked to examine the relevance of the remaining shots in the pool with respect to the given topic

The pooling approach for obtaining relevance judgements assumes that un-judged items are not relevant It can argued that the precision and recall values for systems that did not help building the pool will be underestimated compared to systems that did help This is because the submitted runs from the former could return highly ranked items that are relevant, but have not been judged by any information specialists

Zobel studied TREC text based corpus and found that the quality of the pools affects the quality of the final relevance judgements and the pools for TREC relevance judgements did not show any significant bias against un-judged items [Zobel, 1998] In his test, he chose a run and compared the evaluation difference for the run on two different sets of pool (1) the official TREC pool, (2) a modified pool, which was built by removing items from the official pool that were contributed only by the chosen run For TREC-5 ad hoc data, there was an average improvement 0 5% in using the official pool over the modified pool and the maximum was 3 5%

For TREC-3 ad hoc data, an overall improvement of 2 2% was found Conclusion was drawn for text corpuses that the TREC collection provides a reasonably reliable base for system comparison and the existing TREC relevance judgements can be used to evaluate new retrieval methods

The reliability of the pooling measurement for new video retrieval systems using the TRECVID collection has not been tested by any TREC participants Text retrieval and video retrieval use very different mechanisms The former usually uses semantically rich text to describe concepts and ideas for readers The latter uses not only text (i e ASR transcript or teletext) but also low-level features The ASR text in videos tells us what has been said but very often not what we see on the screen A video retrieval system is encouraged to integrate low-level features to find "difficult" relevant items that can not be found by any text It is likely that a novel video retrieval system returns some "difficult" highly ranked relevant items that are not judged and its performance may be underestimated It is not clear whether the conclusion that "effect of the pool is probably unimportant for the evaluation of video retrieval systems" holds for video retrieval We should be aware that the performance is probably being underestimated

5.3.2 Evaluation Measures Used in TRECVID

Apart from the precision and recall curve, TRECVID uses 3 other evaluation measures (1) precision at different document cut-off levels, (2) average precision by topic, (3) Mean Average Precision (MAP) [Voorhees, 2002]

- Precision at different document cut-off levels

A cut-off level is a rank position that defines a subset of items of a submitted result list for evaluation For instance, a cut-off level at 20 defines the subset as the top 20 items in the ranked list A precision at document cut-off levels is constructed as follow

1 fix the number of items retrieved at several cut-off levels such as top 5, top 10, top 20, top 30 and top 100
2 measure the precision at each of these levels
3 average over all the topics with equal weight among topics

At a single cut-off level, precision reflects the number of relevant documents retrieved Naturally as the cut-off level is increased precision tends to decrease

- Average precision by topic

This is not the average of the precision at recalls, but the mean of the actual precision values obtained after each relevant item is retrieved For instance, a system retrieves three relevant items for a given topic at ranks 2, 6 and 15 The actual precision obtained after each relevant item is retrieved is 0 5, 0 33 and 0 2, respectively Finally, the average precision for the topic is 0 34

Retrieved items with a higher rank will have a higher precision value and vice versa If no relevant items are retrieved, the average precision for the topic is 0 Average precision considers the performance of a system over all relevant items for the topic and a high average precision indicates its ability to retrieve relevant items quickly

- Mean Average Precision (MAP)

Mean Average Precision is a single-valued measure for a system submission that consists of multiple topics It is defined as the mean of the average precision of each topic Comparing the MAPs of two systems, a higher MAP indicates its system performs better for most topics MAP is a summary measure For detailed system comparison, we are referred to other measures described in this section

5.4 Conclusions

Evaluation of information retrieval systems can help us understand the strengths and weaknesses of different systems Precision and recall have been used for many years as the main measurements of system effectiveness Other measures are available for monitoring the specific and overall system performance such as precision at various document cut-offs and mean average precision

The calculation of precision and recall depends on the determination of relevance Relevance is a subjective concept and the judgement of relevance is varied according to users' knowledge and requests at hand An objective view considers the degree to which a document covers the content that is appropriate to the request The objective view is the basic criteria for the relevance judgements in the evaluation environment

The TREC video track provides a large digital video collection as a common baseline for system evaluation A pooling method is used in TRECVID for determining relevant items with respect to a given topic Previous experiments have shown that TREC relevance judgements for text corpuses are reasonably reliable and can be used to evaluate new retrieval methods Since the reliability of the relevance judgements for video collections have not been tested, we should be aware that the performance is probably being underestimated

A complete system evaluation consists of four elements [Salton & McGill, 1983]

1 A detailed system description and its major components The components include query formulation, indexer, retrieval method and output presentation
2 A set of hypotheses to be tested This could be a performance comparison between a particular prototype against an existing retrieval model, or an examination of one or more new system components over components of a baseline system

3 A set of measurable quantities indicating the performance objectives of the system

4 Methods for obtaining and evaluating the data For example, a simple way of measuring the response time is to use a stopwatch

Our retrieval system which we use to evaluate some of the hypotheses in this thesis is an automatic system which is composed of an indexer and a retrieval algorithm The experiments we report later will test the hypothesis that the semantic description of each shot in a collection can be enriched from clusters based on content-based features by using the TRECVID2002 and the TRECVID2003 collections The precision and recall measures are the most widely used and well-understood in experimental information retrieval and are sufficient for the measurement of retrieval effectiveness in our experiments Precision at different document cut-offs, average precision by topic and mean average precision will also be included to explain our experimental results Chapter 6 will study our experiments on the TRECVID2002 collection while Chapter 7 focuses on TRECVID2003

Chapter Six

Experiments on the TRECVID2002 Search Collection

Starting in 2001, the TREC conferences have established a video "track" for researchers who are interested in digital video analysis, indexing and retrieval This track provides a large video collection and uniform evaluation procedures as a baseline for system comparison During the last three years, 2001, 2002 and 2003, the video collection moved from video of 1940s and 1960s documentaries to video of 1998 ABC and CNN news The total number of evaluation hours of MPEG-1 has increased from 11 hours in 2001 to 120 hours in 2003 and is set to increase to 240 hours in 2004 The video track in 2003 covered four tasks (1) shot boundary segmentation, (2) news story segmentation, (3) high-level feature extraction and (4) search

This chapter is concerned with experiments on the manual search task in TRECVID2002 Under this general heading, we will deal with 3 subtopics (1) an overview of the manual search task of TRECVID2002, (2) a review of the TRECVID2002 experimental results of other research groups and (3) our own experimental results in detail The aim of our experiments is to enhance the semantic description of each shot in the collection from clusters based on low/mid level visual features and Automatic Speech Recognition (ASR) texts In the chapter following this, we will turn to the experiments we performed on the TRECVID2003 search collection

Ten systems with different settings were built for the TRECVID2002 collection and nine systems for the TRECVID2003 collection in order to test four evaluation objectives

- Is an aggregated index useful in helping traditional text-only queries?
- Is an aggregated query useful in searching an aggregated index?
- Is there any performance difference between using one best image example and using all image examples in an aggregated query when considering the primitive visual features alone? It would be useful to accept all the example images in a query to avoid the trouble of query selection when users have no knowledge of the search collection
- Is there any performance difference between systems using content-based features alone and systems using concept-based features alone

6.1 The TRECVID 2002 Manual Search Task

The video search task in TRECVID2002 is defined as either fully interactive or manual The fully interactive task supports users' query refinements and other user interaction Most TRECVID2002 systems incorporated relevance feedback and its variations as a query refinement process

The defined video search task for manual searching involves a user formulating a single query for each topic on a once-off basis with no interaction with the video retrieval system, and this query was then used to search the collection A user was presented with a topic description and asked to formulate a query so as to return a ranked list of 100 shots that meet the information need In our work, we are involved in experiments as part of the manual search task and we have manually formulated searches from topics

TRECVID2002 search test collection

The search data for TRECVID2002 contains 176 MPEG-1 videos, 40 12 hours in total The videos are documentaries from the late 1930s to early 1970s Some of the oldest ones were produced to present visual information only without any audio Some of the 'newer' ones show the limitations of early colour video Cartoon features are also used in a number of documentaries

TRECVID2002 topics

There are 25 topics defined in TRECVID2002 Each topic comes with a textual description of the information need along with one or more video clips and/or still images for illustration (see the overview Table 6-1) The layout of the summary table contains

- Text description of the topics
- The number of examples showing the information need
  - Image examples
  - Video examples
- The total number of relevant shots for each of the given topics in the search collection

Examples of the topics were mostly taken from materials from other public resources A few examples come from the search test collection, for instance, the example videos of topic 76 showing "James Chandler"

Table 6-1 Overview of the 25 topics of TRECVID2002 search task

Topic ID	Textual Description of topics	Number of examples		Total Number of relevant shots
		Image	Video	
75	Find shots with Eddie Rickenbacker	2	2	15
76	Find additional shots with James H Chandler		3	47
77	Find pictures of George Washington	1	1	3
78	Find shots with a depiction of Abraham Lincoln	1	1	6
79	Find shots of people spending leisure time at the beach, for example walking, swimming, sunning, playing in the sand Some part of the beach or buildings on it should be visible		4	55
80	Find shots of one or more musicians a man or woman playing a music instrument with instrumental music audible Musician(s) and instrument(s) must be at least partly visible sometime during the shot		2	63
81	Find shots of football players		4	15
82	Find shots of one or more women standing in long dress Dress should be visible at some point		3	18
83	Find shots of the Golden Gate Bridge	5		33
84	Find shots of Price Tower, designed by Frank Lloyd Wright and built in Barlesville, Oklahoma	1		4
85	Find shots containing Washington Square Park's arch in New York City The entire arch should be visible at some point		1	7
86	Find overhead views of cities – downtown and		4	105

	suburbs The viewpoint should be higher than the highest building visible			
87	Find shots of oil fields, rigs, derricks, oil drilling/pumping equipment Shots just of refineries are not desired		1	40
88	Find shots with a map (sketch or graphic) of the continental US		4	72
89	Find shots of a living butterfly	2		10
90	Find more shots with one or more snow-covered mountain peaks or ridges Some sky must be visible behind them		3	75
91	Find shots with one or more parrots	1	1	17
92	Find shots with one or more sailboats, sailing ships, or tall ships – with some sail(s) unfurled	4	2	47
93	Find shots about live beef or dairy cattle, individual cows or bulls, herds of cattle		5	161
94	Find more shots of one or more groups of people, a crowd, walking in an urban environment (for example with streets, traffic, and/or buildings)		3	303
95	Find shots of a nuclear explosion with a mushroom cloud		3	17
96	Find additional shots with one or more US flags flapping		2	31
97	Find more shots with microscopic views of living cells		2	82
98	Find shots with a locomotive (and attached railroad cars if any) approaching the viewer		5	56
99	Find shots of a rocket or missile taking off Simulations are acceptable		2	11

The relevance judgements for each topic are important for the test collection A full set of relevance judgements for each topic on the 14,451 shots would have been infeasible, that could have resulted in over 350,000 judgements The method used in TRECVID2002, known as the pooling method, is to make relevance judgements only on the shots submitted by the different participating systems

To construct such a sample pool for each topic, the top 50 (half) items for each submitted TRECVID2002 run are collected and duplicate items are removed The relevant shots for a given topic are compiled from all relevant items in the pool judged by a NIST assessor

Defined retrieval unit

To compare systems' performance, a commonly agreed retrieval unit was used and a shot instead of a video is defined in TRECVID2002 to serve this purpose Common shot boundary definitions were created by the CLIPS-IMAG group and formatted in

MPEG-7 by ourselves for distribution to all TRECVID2002 participants For each shot we record the start time and the shot duration There are 14,451 shots in total for the complete search collection

Indexing features available in TRECVID2002

Two types of indexing features were marked up in MPEG-7 format and shared among participants These were the ASR transcripts donated by LIMSI [Gauvain et al, 2000] and the output of 10 feature detectors donated by TRECVID2002 groups The ASR transcripts were presented as text annotations corresponding to each of the shots

The feature detectors defined by TRECVID2002 participants included outdoors, indoors, cityscape, landscape, face, people, text overlay, speech, instrument and monologue For each of these the results show the presence or absence of the features within each of the common shots

6.2 A Review of Video Retrieval Techniques Used by TRECVID2002 Participants

There are many ways of describing an information need in video retrieval, textual (query words), visual (image, video), and possibly a combination of both Query words name things, concepts, qualities or ideas, feelings and that alone can tell a great deal about the information need Searching for shots by query words can facilitate the quality of abstraction or concreteness and the level of formality or informality of words By contrast, image/video examples describe things that can be more concrete and specific, things that can be seen, felt and touched They are vivid and emphatic Image/video examples however can not be used to express abstraction such as concepts and ideas

All TRECVID2002 participating systems are designed to handle both query words and visual examples For each video example, its key frame is extracted by different participating systems and is handled in the same way as still image

examples We use the phrase - image examples in the rest of the thesis to refer to both video and image examples

Given more than one image example for illustrating a search topic, there is a slightly different opinion about how many images should be chosen in a query from participants As part of TRECVID2002, groups from IBM, Imperial College, University of Maryland and University of Oulu developed systems that support various ways of selecting one or more images in a query [Adams et al, 2002] [Rautianien et al, 2002] [Wolf, 2002] [Pickering et al, 2002] In particular, the image queries used by IBM were optimised over the feature test collection prior to the search Fusion techniques were applied to combine results from the chosen image examples to produce the final ranked list

However, the Lowlands team suggested that combining results from all image examples for a query is problematic and using certain image examples in a query performs better than the others, particularly if examples are from a similar collection [Westerveld et al, 2002] When the result returned from one of the images is poor, the system performance of the combined result could be degraded Retrieval performance can be improved when the results from all images provide moderate rankings Following this, our experiments were set up to show the difference of performance between using one image and multiple images in a query

TRECVID2002 participants were allowed to use ASR transcripts, donated feature extraction results and their own audio/visual features to build their retrieval systems We classify these into three types of features, each of which is considered at the shot level

- *Spoken text feature*

  This is the ASR transcriptions from either the LIMSI donation [Gauvain et al, 2000] [Barras et al, 2002] or an ASR transcript generated by the individual groups if they have the resources to do this

The donated transcriptions on TRECVID2002 were performed by LIMSI using their broadcast news speech transcription system but no direct error rate evaluation of ASR performance on this video data has yet been reported The word error rate on targeted French audio-visual news is in the 25-30% range [Barras et al, 2002] and on the English broadcast-news for the TREC-9 Spoken Document Retrieval is around 20% [Gauvain et al, 2000] Even though the spoken text feature suffers a certain level of word error rate, good retrieval performance can still be maintained [Barras et al, 2002]

To get an estimate of the word error rate for the ASR data from TRECVID2002, three videos were selected from the TRECVID2002 search collection and we manually calculated the approximate word error rate, and these results are shown in Table 6-2

Table 6-2 Summary of the approximate word error rate of the ASR transcriptions of the TRECVID2002 search collection

Video ID	Video Name	The length of the video	Length examined	Approximate Word Error Rate
104	Mountain skywalker	28 min	12 min 18 sec	11 16%
96	Men, steel and earthquakes	27 min	8 min 29sec	11 52%
60	Earthquake	26 min	10 min	39 2%

The word error rate of video 60 is much higher than the others because that video is a typical type of expository documentary which is mixed with a great number of re-enactments during the video The expository documentary is considered to be the oldest form of documentary dating back to 1930s It often employs a narrator to interpret the visual pictures for the audience In cases where filmmakers cover events after their occurrence and want to address some set of events that happened, they explicitly reconstruct the original action in a studio environment They have actors who are not narrators playing the roles of real people

To take an example from video 60 about "Earthquake", one of the scenes would be about the warning of a potential earthquake to the public To heighten the tension of the situation the filmmakers reconstructed a number

of phone conversations from a weather observatory to authorities telling of the disaster During the re-enactments different actors but not the narrator were involved and the high-speed rate of the conversations made it easier to convey the impression of the forthcoming earthquake These may be the causes of the low word error rate in this and similar videos

- *Concept – based features*

These are the feature donations by some of the TREVID2002 participants which are extracted based on pre-defined and agreed concept labels These concept labels include indoors, outdoors, face, people, cityscape, landscape, speech, instrumental sound, monologue and text-overlay Each shot is associated with a concept label with a confidence value ranging from 0 to 1 Most TRECVID2002 groups who used these donated features as part of searching needed to decide cut-off points to turn their confidence value into a binary value so as to show whether the concept is found in a given shot [Rautianien et al, 2002]

Concept labels can be seen as ontological categories Ontological categories are vocabulary for describing things that exist and provide knowledge of the physical world The selection of these categories (or concepts) determines things that can be represented within a video

If the selected concept is too general such as "speech" and "instrumental sound", it might be easy to detect but might not be useful for retrieval due to its frequent appearances in videos If the concept is too specific such as "Bill Clinton" and "Mercedes-Benz logo", it would be useful but the detection of its presence/absence in a shot remains a challenge Even though there are some content-based features that can be modelled as semantic concepts (i e things), there are a number of concepts for which the creation of a model is impossible (i e abstractions, feelings and ideas)

Also, any incompleteness in the ontological categories restricts the generality and completeness of representations of video content Composing a query using concept labels therefore became problematic because there are so many

things that can be included in a topic but there are so few concept labels we can choose from

- *Content – based features*

These are the low-level features such as colour, texture, shape and motion Features within this type can be used to model concept-based features The problem of using content–based features for video retrieval is that the features themselves lack semantic information Conventional content–based techniques can help find limited things, for example that have similar colour and shape, provided that query image examples are similar to the search collection m terms of the image quality Content-based features were found useful m the relevance feedback process rather than in the query initialisation process [Adams et al, 2002] and this is because the images selected for relevance feedback are from within the search collection

The detection of concept-based features introduces another way of doing video retrieval The group from University of Maryland treated these features in the same way as the spoken text and content-based features [Wolf et al, 2002] They generated a concept feature vector for each shot and each cell of the vector is a binary value of a concept-based feature, illustrating whether a concept-based feature is present in the shot or not If there are 10 features, the vector will have a length of 10 A similarity formula was given for comparing concept-based features

The team from the University of Oulu decided to use concept-based features as binary filters in order to create subsets of a concept or combinations of concepts prior to retrieval [Rautiainen et al, 2002] In short, concept-based features worked better when a query topic was a close match to the concept label, for example, concept "face" may create a good filter for topics looking for a person like "Eddie Rickenbacker" and "James Chandler"

Table 6-3 below summarises the type of features used in the manual search task by 6 TRECVID2002 participants Their video retrieval systems were developed without knowledge of search test data These features include the spoken document

feature, concept-based and content-based features Also noted in Table 6-3 is the query formulation techniques used if any

Table 6-3 Summary of video retrieval approaches used in manual search task by TRECVID2002 participants

Group	Spoken document features	Concept-based features	Content-based features	Query formulation techniques
Fudan University	Vector space model	None	Camera motion, face dectection, Text-overlay detection, Speaker Recognition	Multiple images in a query
IBM Research	Vector space model	All 10 features plus their own concept-based features	Colour histogram, colour correlogram, edge orientation histogram and etc	Multiple images in a query
Lowlands Team	Language model	None	A mixture of Gaussian models of DCT coefficients in the YCbCr colour space	• One best image example in a query, chosen based on the criteria that it is comparable to the search collection • Images from Google search engine were also used in a query
Imperial College	Vector space model	None	Colour histogram, colour correlogram, edge orientation histogram and etc	Multiple images in a query
University of Maryland	Vector space model	All 10 features were used as a query vector	Temporal colour correlograms	Multiple images in a query
University of Oulu	Vector space model	All 10 features were used as a filter	Colour, Motion, and audio	Multiple images in a query

An increase in the types of detectable features provides a video retrieval system with more flexibility but also more complications In initialising a query, it allows the users flexibility to choose the features they wish for each query, but meanwhile may create a difficulty for them to build the optimal queries Users often specified multiple features in a query, and switch search strategies among features when searching for a topic

As a result, various methods of combining results from independent searches based on different features were emphasised by most TRECVID2002 participants though no best solution to how to combine these was found Attempts to combine

were done using a simple weighted sum schema to combine the results (the teams from IBM, the Lowlands team and Imperial College)

Our interest in these search experiments in TRECVID2002 are drawn to the role of the spoken-text, concept-based and content-based features extracted from videos on their own, and by combination, whether these could possibly work together to improve video retrieval

Table 6-4 below summarises the performance of the manual search by the TRECVID2002 participants There were 19 runs in total submitted by 6 groups who were interested in developing a system where the search was performed without any knowledge of the search collection A system specification is given in the table to show what types of features were used for a given run It should be noted that not all submitted runs ran all 25 topics, i e for runs submitted by the group from Fudan University

Later work by Thijs Westerveld from CWI [Westerveld et al, 2003] pointed out that combining textual and visual features can improve retrieval effectiveness when text-only runs and image-only runs yield reasonable results, respectively In other words, retrieval performance for the combined retrieval can be degraded when one type of information produces good results while the other gives poor results

Table 6-4 Summary of performance of the manual search task by TRECVID2002 participants

Group	Submitted Run Codes	Mean Average Precision	Total topics submitted	Total shots returned	Total relevant shots	Total relevant shots returned	System Specification
Fudan	Fudan_ Sys1	0 0632	13	1201	962	33	Content-based only
	Fudan _Sys2	0 0053	5	500	383	12	Their own Concept-based
	Fudan _Sys3	0 0717	10	1000	765	27	Donated Concept-based
	Fudan _Sys4	0 0804	20	1353	1140	125	ASR + Concept-based
IBM	IBM-1	0 0059	25	2498	1445	64	Content-based + Concept-based
	IBM-2	0 1357	25	2500	1445	256	ASR
	IBM-3	0 0926	25	2497	1445	229	ASR + Content-based + Concept-based

Imperial	ICMKM-2	0 06	25	2491	1445	138	ASR + Content-based
	ICMKM-3	0 043	25	2492	1445	133	ASR + Content-based + Concept-based
	ICMKM-4	0 0568	25	2491	1445	132	ASR + Content-based
Lowlands	LL10_T	0 0917	25	2456	1445	181	ASR
	LL10_Tiac	0 0016	25	2497	1445	30	ASR + Content-based
	LL10_Tisc	0 0022	25	2499	1445	35	ASR + Content-based
	LL10_TIScG	0 0038	25	2499	1445	26	ASR + Content-based
Maryland	UMDMNAqt	0 0256	25	1862	1445	91	Content-based + Concept-based
	UMDMqtrec	0 0587	25	2364	1445	171	ASR + Content-based + Concept-based
Oulu	UnivO_MT1	0 0333	25	1330	1445	89	Content-based
	UnivO_MT2	0 0194	25	1066	1445	57	Content-based + Concept-based
	UnivO_MT3	0 0061	23	690	1413	43	ASR + Content-based

The experimental results from TRECVID2002 show that using query words alone is more effective for video retrieval than using image examples only or using a combination of both [Smeaton & Over, 2002] In other words, systems built based on the spoken-text feature alone generally outperformed systems that incorporated selected combinations of three types of features, i e spoken-text, content-based and concept-based features

The results obtained by the TRECVID2002 participants in the manual search are counter-intuitive and surprising and were a set-back for progressing video IR One of the reasons for this could be that reported accuracy of detection of the concept-based features donated by TRECVID2002 groups is poor

It was also found that a system that supports the spoken-text feature and any other types of features worked better than a system that uses content-based / concept-based alone However, there is a discrepancy to this finding in the manual search results provided by the University of Oulu (see Table 6-4 above) The introduction of the spoken-text feature into a content-based video retrieval system caused a big drop in Mean Average Precision from 0 033 to 0 006 No specific reason to this was given in the original work

Figure 6-1 below shows the average precision by topic for the 19 runs in the manual search task in TRECVID2002. The mean of the top half of results for topic 76,77 and 84 stand out and show these topics to be high performing ones.

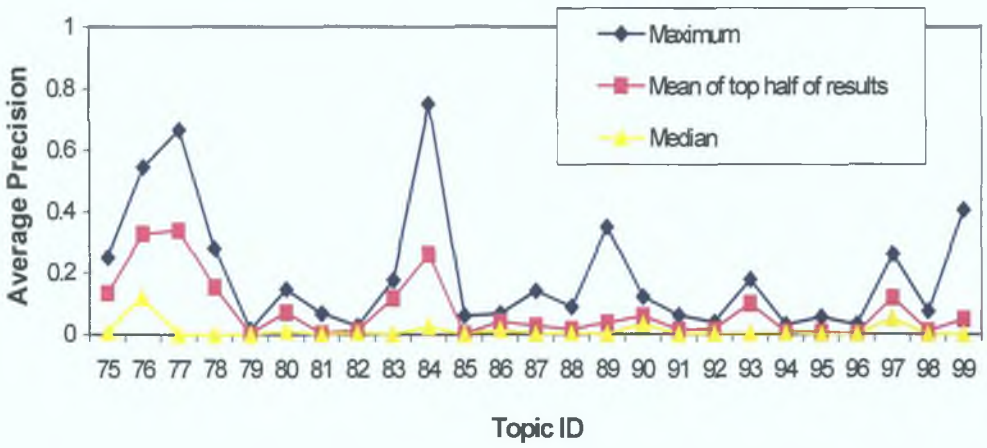


Figure 6-1: Manual search: average precision per topic

The TRECVID2002 results are not meant to be conclusive and there are several arguable flaws with the appropriateness of the data set used, the feature extraction accuracy, the topic definition method and the depth of the pool of relevance judgements.

TRECVID2002 represents one of the few available video IR collections and the experiments in the manual search task do illustrate a number of least to most favourable options in considering visual features during search. The spoken-text feature is thought to be a reasonably accurate property of videos. The lack of semantic information is one of main shortcoming of using content-based features for video retrieval while the concept-based feature detection task in TRECVID2002 did try to establish a link between content-based features and semantic concepts but still requires a great deal of research to improve the performance of the subsequent search.

6.3 Experimental Settings

Having studied the TRECVID2002 experiments, we understand the difficulties of applying concept-based / content-based features in video retrieval To obtain better retrieval performance, spoken-text features must be employed The experiments in this thesis were set out to study whether applying content-based features or concept-based features in conjunction with spoken-text feature will give better performance

Also of interest to us is the system performance difference between one image example and all examples in a TRECVID2002 query when content-based features are considered The reason for choosing one best example in a query was addressed by [Westerveld02] as reported in section 6 2 Our proposed video retrieval model hopes to use all the example images in a query to avoid the trouble of query example selection when users have no knowledge of the search collection such as the quality of pictures

The experimental settings for our retrieval experiments are described in section 6 3 1 It details the construction of a baseline model and various modified systems for comparison Section 6 3 2 will illustrate a run-through example of how our systems work Finally, our experimental results will be given in section 6 3 3 in detail

6.3.1 Experimental Settings

Our baseline system

A conventional video retrieval model using the spoken-text feature alone was implemented as a baseline for evaluating the proposed model The indexing model used is a Vector Space model adopting the Term Frequency Inverse Document Frequency (TF*IDF) weighting algorithm A shot is considered as a retrieval unit The ASR transcript that belongs to the shot was solely contributing to forming a term weight vector No ASR transcripts coming from temporal neighbours of the shot were included into the formation of the term weight vector of the shot The

conventional video retrieval model was constructed as follows in terms of indexing and retrieval method

- Indexing
  (1) Remove stop words,
  (2) Stem words into their basic forms (tokens),
  (3) Calculate the frequency of occurrences of each token W_j in a given shot S_i, noted as Term Frequency, TF_{ij},
  (4) Calculate the number of unique shots in the collection that contain token W_j, noted as Document Frequency, DF_j,
  (5) Calculate the number of shots in the collection, noted as n,
  (6) Obtain the TF*IDF term weight SW_{ij} for each token in a given shot based on formula 6-1

$$SW_{ij} = TF_{ij} * [Log_2(n) - Log_2(DF_j) + 1] \quad (6\text{-}1)$$

- Retrieval
  (1) Remove stop words from a query,
  (2) Stem query words into their basic forms called query tokens,
  (3) Construct a query vector with length n, where n is the total number of unique tokens found in the search collection Each cell of the query vector is a binary value, "1" indicating the presence of a query token in the search collection and "0" in the absence of the query token, noted as QW_j,
  (4) Calculate the shot similarity Sim_i of shot S_i to the given text query based on Formula 6-2

$$Sim_i = \sum_j QW_j * SW_{ij} \quad (6\text{-}2)$$

  (5) Rank shots based on decreasing similarity value Sim_i in descending order and return the top 100 most similar shots as a submission for evaluation

Query preparation

Our experiments belong to the manual search category of TRECVID2002, emphasising that the query from a given topic is formulated on a once-off basis and so formulating an optimal query is required though no optimal query is possible in our experiments Instead, the query for each topic was formulated based on three types of features We name them as follows

- *Original text query* – Query terms in this category were to come from the topic descriptions, see Table 6-5 below No other words outside the topic descriptions were included

Table 6-5 Selected text and manually created concept-based query vector by topic in TRECVID2002

Topic ID	Text query	Concept-based query									
		Outdoors	Indoors	Textover	Peole	Landscape	Cityscape	Monologue	Face	Sound	Speech
75	Eddie Rickenbacker	0.5	0.5	0	0.5	0.5	0.5	0.5	1	0.5	0.5
76	James Chandler	0.5	0.5	0	0.5	0.5	0.5	0.5	1	0.5	0.5
77	George Washington	0	1	0	0.5	0.5	0.5	0.5	1	0	1
78	Abraham Lincoln	0	1	0	0.5	0.5	0.5	0.5	1	0	1
79	People spending leisure time at the beach, walking, swimming, sunning, playing sand	1	0	0	1	1	0	0	0	0.5	0.5
80	Musicians man woman playing a instrument with instrumental music	0	1	0	1	0	1	0	1	1	0
81	Football players	1	0	0	1	0.5	0	0	0	0.5	0.5
82	Women standing in long dress	0.5	1	0	0.5	0.5	0.5	0	0	0.5	0.5
83	Golden Gate Bridge	1	0	0	0	1	0	0	0	0.5	0.5
84	Price Tower, designed by Frank Lloyd Wright and built in Barlesville, Oklahoma	1	0	0	0	0	1	0	0	0.5	0.5
85	Washington Square Park's arch in New York City	1	0	0	0.5	0	1	0	0	0.5	0.5
86	Overhead views cities downtown and suburbs	1	0	0	0	0	1	0	0	0.5	0.5
87	Oil fields, rigs, derricks, oil drilling pumping equipment	1	0	0	0	0.5	0	0	0	0.5	0.5
88	Map continental US	0	0.5	0	0	0	0	0	0	0.5	0.5
89	Living butterfly	1	0.5	0	0	0.5	0.5	0	0	0.5	0.5
90	Snow covered mountain peaks or ridges	1	0	0	0	1	0	0	0	0.5	0.5
91	Parrots	1	0.5	0	0	0.5	0.5	0	0	0.5	0.5
92	Sailboats, sailing ships, or tall ships with some sail unfurled	1	0	0	0	1	0	0	0	0.5	0.5
93	Live beef or dairy cattle, individual cows or bulls, herds	1	0	0	0	1	0	0	0	0.5	0.5

	of cattle												
94	Groups people, crowd, walking urban environment streets, traffic buildings		0	0		0		0	0	0 5	0 5
95	Nuclear explosion mushroom cloud		0	0	0	0 5	0	0	0	0	0 5
96	US flags flapping		0	0 5	0	0		0	0	0 5	0 5
97	Microscopic views living cells	0		0	0	0	0	0	0	0 5	0 5
98	Locomotive railroad cars approaching		0	0 5	0	0 5	0	0	0	0 5	0 5
99	Rocket missile taking off Simulations		0	0	0	0	0	0	0	0 5	0 5

- *Concept-based query* – Given a topic, we manually assigned a value from {0, 0 5, 1} to each concept-based feature These values together form a query vector We called this a concept-based query vector which is also seen as a non-text query (see Table 6-5 above)

  - Value "*0*" the absence of the given feature is of no concern for a given topic To take an example of topic 88 "finding shots with a map of the continental US", we assigned *0* to feature "Face" and "People"

  - Value "*0 5*" the presence or absence of the given feature is thought not to affect a given topic Feature "text-overlay" and "instrumental sound" were often given 0 5 because the relevant shots do not care about their presence or absence within the shots

  - Value "*1*" the presence of the given feature is considered to be important, for instance, in topic 75 "find shots of Eddie Rickenbacker", the feature "face" must be present with a value *1*

Concept-based query vectors were constructed subjectively Certain topics created the same query vectors This is partly because of the generality of the concept features provided by TRECVID2002 For instance, topic 92 and 93 are to find "sailboat" and "dairy cattle" respectively The features "Outdoors" and "landscape" were switched on for both topics and so both have the same concept-based query vectors

- *Content-based query* – A content-based query is also a non-text query and was automatically computed for each image example provided by TRECVID2002. This type of queries are objective since no human is involved in creating them. Two colour-based features and one texture-based feature were used to represent each image example. The definition of each feature is given in the MPEG-7 representation of visual features (see section 2.2).
  - 9 region * dominant colour using the RGB colour space. The number of dominant colours for each RGB colour component is 2.
  - 4 region * 16 bin scalable colour (i.e. colour histogram).
  - Global 80 bin edge component histogram.

Indexing preparation

Three types of features are considered separately to build the index. The conventional TF*IDF indexing method is adapted to the spoken-text feature at the shot level, see Formula 6-1. Since the process of creating an index for concept-based features and content-based features is the same, only the process for content-based features are summarised below.

(1) Apply the K-means shot clustering algorithm to obtain clusters for each video.

(2) Assign meanings to each cluster using a modified TF*IDF algorithm. TC_{kj} is the frequency of occurrences of each token W_j in a given cluster C_k. CF_j gives the number of unique shots in the collection that contain token W_j. nc indicates the total number of clusters created in the whole search collection. The weight of token W_j in a given cluster C_k is given in Formula 6-3.

$$CW_{kj}^{content} = TC_{kj} * [Log_2(nc) - Log_2(CF_j) + 1] \qquad (6\text{-}3)$$

(3) Derive the meanings of a shot based on its corresponding cluster meanings. The distance between a given shot and its corresponding cluster centre is noted as $Dist_{ik}$. The weight of token W_j in a given

shot S_i inferred from the associated cluster C_k , $DerivedSW_{ij}^{content}$, is shown in Formula 6-4

$$DerivedSW_{ij}^{content} = (1 - Dist_{ik}) * CW_{kj}^{content} \qquad (6\text{-}4)$$

(4) Aggregate the three term vector weights of a given shot to produce a single final term weight index for video retrieval according to the methods of formulating a query

Three separate sets of term weight indexes are created in the indexing stage

- SW_{ij}, obtained from the conventional TF*IDF algorithm based on the spoken-text feature
- $DerivedSW_{ij}^{content}$, derived from content-based features
- $DerivedSW_{ij}^{concept}$, derived from concept-based features

Query Preparation

Given three different types of queries for each TRECVID2002 topic (i e the original text, content-based and concept-based query), an attempt was made to join them together into a single form of query, namely a text query Content-based and concept-based queries are mapped onto a text query by two steps, respectively Listed below are the steps showing how to map a content-based query onto a text query

(1) Find the *topK* most similar clusters C_m ($1 \leq m \leq K$) to a content-based query $Q^{content}$ *topK* is the number of the most similar clusters The clusters used are obtained based on content-based features The similarity between a cluster and the content-based query is determined by the distance between the cluster centroid and the query, noted as $D_m(Q^{content}, C_m)$

(2) Take the term vector of the *topK* chosen clusters and add them together to form a single query term vector based on the normalised inverse distance $NID_m^{content}$ The weight of query term T_j, noted as $QW_j^{content}$, is given in Formula 6-5 and 6-6, where $CW_{mj}^{content}$ is the weight of token W_j in a given cluster C_m

$$QW_j^{content} = \sum_m CW_{mj}^{content} * NID_m(Q^{content}, C_m) \quad (6\text{-}5)$$

$$NID_m = 1 - \frac{D_m}{\sum_{m=1}^{K} D_m} \quad (6\text{-}6)$$

When multiple image examples including the key frames of video examples are used in a query, each image example $Image_a$ ($1 \le a \le S$, S is the number of image examples) is first mapped to a term query vector $QW^{content\ image}{}_a$ based on content-based features as described above in Formula 6-5 We then combine these a term query vectors to form a single term query vectors for all the image examples using an aggregation technique as shown below in Formula 6-7, where Agg() is an aggregation function

$$QW^{content} = Agg(QW^{content-image}{}_1, \ , QW^{content-image}{}_S) \quad (6\text{-}7)$$

The aggregation technique is used in structured document retrieval for combining scores of various items according to an associated importance value which is specified by a linguistic quantifier (see section 3 5 for details) We use the quantifier "at least ½" in our work to emphasise high scores by placing more importance values to them

Three types of query have been transformed into a term weight vector format

- QW_j, obtained directly from manually selected query words see Table 6-6
- $QW_j^{content}$, derived from content-based query
- $QW_j^{concept}$, derived from concept-based query

Evaluation of systems

We designed nine systems for evaluation based on the selection of indexes and query term vectors as shown in Table 6-6 System "Sys1" is our baseline system which implements the conventional Vector Space retrieval model The rest of the systems incorporate the aggregation process in both the indexing and retrieval process accordingly

Table 6-6 Evaluation system design for TRECVID2002

		Methods of formulating a query for a given topic					
		QW_j	$QW_j^{content}$	$QW_j^{concept}$	$QW_j + QW_j^{content}$	$QW + QW_j^{concept}$	$QW + QW_j^{content} + QW_j^{concept}$
Scope of index aggregation	SW_{ij}	Sys1					
	$SW_{ij} + DerivedSW_{ij}^{content}$	Sys2	Sys3		Sys4		
	$SW_{ij} + DerivedSW_{ij}^{concept}$	Sys5		Sys6		Sys7	
	$SW_{ij} + DerivedSW_{ij}^{content} + DerivedSW_{ij}^{concept}$	Sys8					Sys9

To explain this, we take an example of "Sys4" The process of returning the top 100 relevant shots is described as follows

(1) Aggregate index SW_{ij} and $DerivedSW_{ij}^{content}$, where Agg() is an aggregation function The aggregated weight of token W_j of a given $Shot_i$ is noted as $AggSW_{ij}$

$$AggSW_{ij} = Agg(SW_{ij}, DerivedSW_{ij}^{content}) \quad (6\text{-}8)$$

(2) Aggregate query term vectors QW_j and $QW_j^{content}$ The aggregated weight of query token W_j is given as $AggQW_j$

$$AggQW_j = Agg(QW_j, QW_j^{content}) \quad (6\text{-}9)$$

(3) Calculate the shot similarity Sim_i to the given aggregated text query based on Formula 6-10

$$Sim_i = \sum_j AggQW_j * AggSW_{ij} \quad (6\text{-}10)$$

(4) Rank shots based on the similarity value Sim_i in order and return the top 100 most similar shots as a submission for evaluation

Variable Setting

Unlike System 1, Systems 2 to 9 depends on three variables and the following experiments on the variables were carried out, respectively

- *The window size used in the K-means shot clustering and the threshold*

  A video consists of a sequence of shots It is reasonable to consider the temporal order of shots during clustering If any of the candidates, the adjacent shots of a given shot S_i, are similar enough to the shot, it should be put into the same cluster as S_i The window size *WIN* was varied from 0 to 3 Given a window size of 2, for instance, the next 2 shots of a shot in the sequence would be the candidates

  A threshold T is required to determine the "similar enough" criteria But there is no direct way of setting the threshold without the knowledge of distribution of the dissimilarity values between shots An indirect method to assign the threshold is described for each video as below

  (1) Given a window size *WIN*, calculate the dissimilarity values between a shot and its candidate shots If a given video contains N shots, there would be $WIN * N$ dissimilarity values in total

  (2) Sort the $WIN * N$ dissimilarity values in ascending order Shots are similar enough if their dissimilarity values are within the range set by the top xT percent of all sorted dissimilarity values The dissimilarity value in position $WIN*N*xT$ is the threshold for the video

  The variable xT is chosen to be 0 1, 0 15 or 0 2 The chosen value pairs for variable *WIN* and xT are listed in Table 6-7

Table 6-7 Chosen value pairs for variable WIN and xT

WIN	0	1	1	1	2	2	2	3	3	3
xT	0	0 1	0 15	0 2	0 1	0 15	0 2	0 1	0 15	0 2

- *The top K most similar clusters to a non-text query*

  This variable considers the number of most similar clusters chosen to map a non-text query onto a query term vector In other words, it accounts for the amount of query terms that are inferred to describe the meaning of the non-text query The non-text query refers to a content-based query extracted from

an image example or a concept-based query manually constructed by us The number of *topK* is chosen to be 1, 3, 5, 7 or 9

- *The proportion of weight PF given to an original text query when creating a combined query*

  This variable considers weights given to original text query terms An outcome of using derived term weights of non-text queries is the possibility of bringing more unrelated terms to a given topic It thus loosens the specificity of the topic thereafter degrading the system performance It is known that original text queries play an important role in video retrieval Prior to aggregating different types of queries, it is suitable to assign more important values to the original query terms than the derived terms Our attempt is to maintain the extent of topic specificity but also to include additional information from the derived terms

  But how much more weight we should assign to the original query terms is not clear Variable *PF* is introduced to measure the degree of impact of original text queries on the retrieval performance after combining non-text queries The original query term weight QW_i is given in Formula 6-11, where W is a binary value indicating the presence or absence of a term, *max()* is a function to find the maximum value of the derived term weights $QW_i^{content}$ The derived term weight $QW_i^{content}$ could be obtained by either content-based or concept-based queries The weight *PF* is chosen to be 4, 8, 12, 16, 20 or 24

$$QW_i = W + \max_i (QW_i^{content}) * PF \qquad (6\text{-}11)$$

A major distinction between our systems and the TRECVID2002 participants' manual search systems is that our systems attempt to merge the different types of features in the indexing and query formulating stage and no combination of ranked results is required (see Figure 6-2 below), whereas other TRECVID2002 systems regard each types of feature query as separate searches and utilise a sum weighted schema to combine the ranked results from the different searches (see Figure 6-3 below)

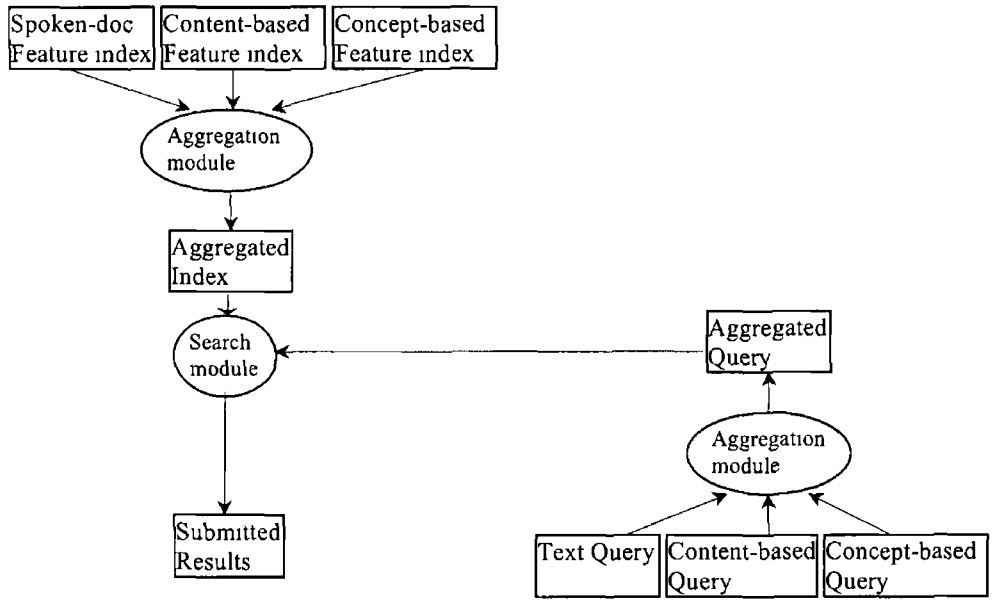


Figure 6-2 Data flow diagram for our systems

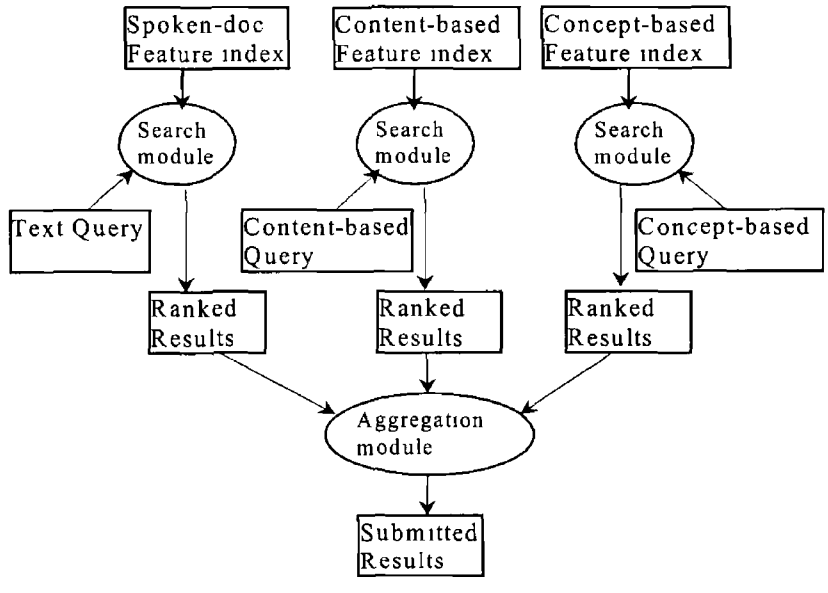


Figure 6-3 Data flow diagram for a typical TRECVID2002 participants' system

Evaluation objectives

The following summarises the objectives of our system evaluations

- Is an aggregated index created from concept-based and content-based features in conjunction with the spoken-text feature useful in helping text only query? To answer the question, Sys2, Sys5 and Sys8 will be studied

- Is an aggregated query derived from non-text queries along with the original text query useful in video retrieval? We pair up Sys2 and Sys4, Sys5 and Sys7 as well as Sys8 and Sys9 for further study
- In considering content-based queries alone, is there any performance difference between using one best image example and using all image examples in a query? Experiments will be carried out within Sys4
- Sys3 and Sys6 are additional systems showing the performance difference between systems using content-based features alone and systems using concept-based features alone

6.3.2 A Run - Through Example of How our System Works

We concentrate on examples of how the system works in this section The examples are given for 4 main processes

(1) Index preparation
- Clustering shots using K-means
- Assigning meanings to the created clusters
- Inferring shot term weights from cluster meanings

(2) Query preparation
- Mapping a non-text query onto a term vector query

Clustering shots using K-means

A k-means clustering algorithm is used to group shots that are visually similar together based on the given content-based or concept-based features A sliding window is applied in the clustering algorithm because of the temporal characteristics of videos The cluster centres are randomly initialised The modification of cluster centre will not stop until the members in each cluster remain the same after two iterations The maximum number of iterations is set to be 300 to prevent infinite loops For the detail of shot clustering using content-based features, readers are referred to section 4 2 1

Here, we describe how to apply concept-based features to shot clustering. Given any two shots within the search collection, noted as $Shot_i$ and $Shot_j$, each shot has its corresponding concept-based feature vector $F^{concept}$ with a length of m. In the TRECVID2002 search collection, m equals to 10 and each cell of vector $F^{concept}$ has a confidence value within a range between 0 and 1. The distance $Dist_{concept}$ between $Shot_i$ and $Shot_j$ considering concept-based features is calculated based on Manhattan distance:

$$D_{concept}(Shot_i, Shot_j) = \sum_{p=0}^{m} | f^{concept}{}_p - f^{concept}{}_p | \qquad (6\text{-}12)$$

We run the clustering algorithm for different initial partitions and different K values from a set of integers. The best partition was selected as the one that has the minimum cluster separation measure [Jain & Dubes, 1988] [Ngo et al, 2001] (also see section 4.2.1).

Table 6-8 below gives the distribution of clusters created based on content-based features among a sample of the 20 videos from TRECVID2002 for different variable *WIN* (i.e. the sliding window size) and *xT* (i.e. a variable for obtaining the threshold for deciding whether two adjacent shots are similar enough) settings. Figure 6-4 below shows the distribution in a histogram chart for better visualisation. The last row of the table is the total number of shots in the 176 videos, along with the total number of clusters for each variable setting. It can be seen that the number of clusters generated for each video was not significantly affected by the *WIN* and *xT* settings and it stays almost the same for each pair of values. Results also show that the more shots a video contains, the more clusters can be produced.

Table 6-8: The distribution of clusters created based on content-based features among a sample of the 20 TRECVID2002 videos

Video ID	The total number of shots per video	The number of clusters per video									
		WIN=0	WIN=1 xT=0.1	WIN=1 xT=0.15	WIN=1 xT=0.2	WIN=2 xT=0.1	WIN=2 xT=0.15	WIN=2 xT=0.2	WIN=3 xT=0.1	WIN=3 xT=0.15	WIN=3 xT=0.2
4	105	15	13	14	14	13	13	10	12	13	10
5	74	8	11	8	7	8	8	9	9	8	9
29	49	4	5	3	5	7	4	5	3	3	5
38	65	8	7	6	8	9	8	7	10	10	8

50	14	2	2	2	2	2	3	2	2	2	2
75	207	25	24	25	23	26	25	24	28	26	27
81	101	13	9	9	14	10	10	11	9	15	13
82	38	6	5	4	4	7	4	6	8	6	5
92	56	9	6	9	8	8	8	5	10	8	8
99	165	23	20	20	24	18	22	21	24	22	18
103	62	9	9	9	7	9	9	8	11	9	8
104	254	29	28	28	35	28	35	30	29	31	31
118	59	9	8	5	9	8	7	9	7	8	7
119	176	20	20	22	20	23	26	22	22	21	21
124	63	9	7	5	5	8	8	6	7	6	6
140	57	6	8	7	7	9	9	7	7	7	7
151	102	14	11	14	10	13	11	16	14	14	12
157	28	3	2	3	2	2	3	3	3	3	3
160	9	1	1	1	1	1	1	1	1	1	1
164	75	7	11	9	11	9	10	9	7	8	8
Total	**14451**	**1737**	**1736**	**1741**	**1750**	**1731**	**1760**	**1731**	**1723**	**1751**	**1741**

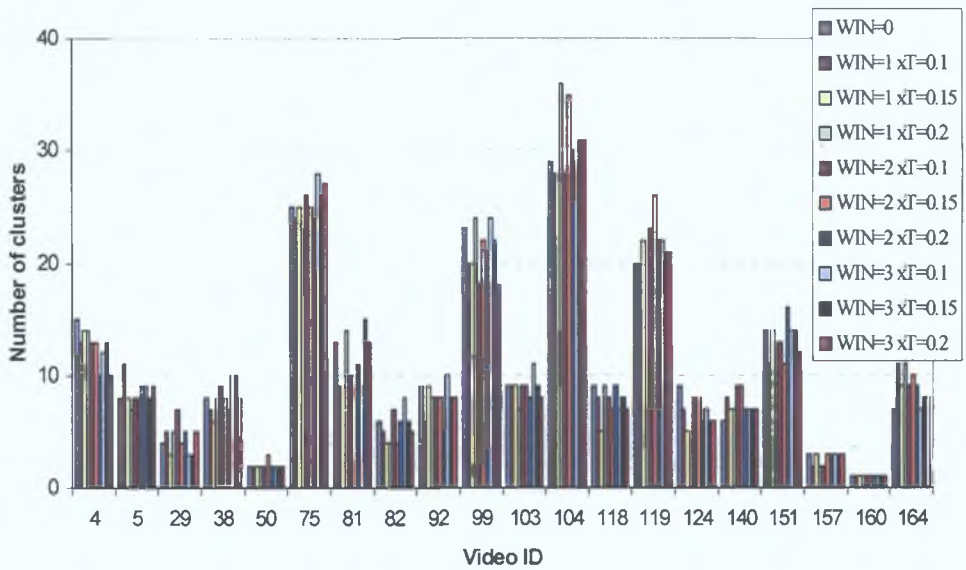


Figure 6-4: The distribution of clusters based on content-based features among the 20 TRECVID2002 videos

Figures 6-5 and 6-6 below show an example of the clusters for video 133 and 158. The clusters were created based on concept-based features. The distance between each shot to its cluster centre is given under its key frame. It can be seen that concept-based features help clustering shots that are similar visually. To take an example of video 133, most shots in cluster133_18 present the feature of "face" and "indoor".

Figure 6-5: Cluster 133_18 generated by using concept-based features

Figure 6-6: An example of a cluster for video 158 generated by using concept-based features

It is difficult for cluster158_5 of video 158 to find its feature concept, in which no people or indoor/outdoor scenes defined by the TRECVID2002 features can be detected. Cluster158_5 groups shots that are mainly showing a living cell. In our work no threshold was set for determining the absence or presence of any concept features according to the confidence value given by feature donations prior to clustering. In other words, we used the confidence that comes from the donations directly to build concept-based feature vector. The confidence value are within a range of [0, 1]. The higher the value is, the more positive is the presence of a concept feature within a given shot. Shots within cluster158_5 are grouped together because of their associated low confidence to all concepts which indicate the absence of most concepts.

Concept-based features are distinguished from content-based features by their level of semantics. It has been pointed out earlier in section 6.3.1 that concept-based features are bound to some general concept labels although the number of concept labels that have been developed is limited and its accuracy of assigning the concepts to shots is problematic. The basic intention of K-means clustering using concept-based features was seen to be possible – to group shots that are perceptually similar in terms of semantics. However if there are many shots within a video with which no concept features could be associated, shots in the same cluster may not present any conceptual similarity (see Figure 6-7 below).

Video 7 - "Aluminium on the March (Part II)"

Figure 6-7: Cluster7_8 generated by using concept-based features

Content-based features for shot clustering groups shots that are similar in terms of colour and edges in such a way as to reduce the disadvantages of unavailable concept-based features. Content-based features are the most basic visual features and they are the building blocks for creating concept-based features. Cluster133_14 and Cluster158_6 are found to have shots that are very similar with respect to colour and edges (Figure 6-8 below).

Figure 6-8: Cluster133_ 14 and cluster158_6 created by using content-based features

Clustering yields some groupings within which shots are not similar in terms of the chosen features. We call these dissimilar clusters. Table 6-9 below summarises the number of dissimilar and similar clusters from 20 videos for concept-based and content-based features, respectively, as determined from a manual inspection. The judgement about whether a cluster is a similar cluster or not is based on the similarity of the features chosen for the clustering process rather than purely on semantic information. The process was done subjectively by visualising the shots from each cluster. The results have shown that clustering based on content-based features gives more similar clusters in terms of low-level features than clustering by using concept-based features. If more concepts are available, the clustering results of concept-based features would improve.

Table 6-9 Summary of clustering results of 20 videos for content-based and concept-based features

Video ID	Content-based similar clusters		Content-based dissimilar clusters		Concept-based similar clusters		Concept-based dissimilar clusters	
	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage
3	7	63%	4	36%	5	42%	7	58%
8	10	67%	5	33%	13	87%	2	13%
22	13	85%	2	15%	10	71%	4	29%
41	9	89%	1	11%	12	85%	2	15%
47	7	70%	3	30%	6	54%	5	46%
54	8	80%	2	20%	6	67%	3	33%
59	7	87%	1	13%	10	100%	0	0%
64	15	88%	2	12%	6	50%	6	50%
73	11	73%	4	27%	7	53%	6	47%
79	5	83%	1	17%	5	71%	2	29%
85	7	64%	4	36%	6	67%	3	33%
93	10	71%	4	29%	11	69%	5	31%
100	7	70%	3	30%	5	50%	5	50%
106	10	90%	1	10%	6	67%	3	33%
114	8	72%	3	28%	6	47%	7	53%
123	12	75%	4	25%	8	67%	4	33%
133	9	75%	3	25%	12	80%	3	20%
141	7	87%	1	13%	5	62%	3	38%
158	10	83%	2	17%	8	67%	4	33%
162	8	72%	3	28%	7	58%	5	42%
Average		**77%**		**23%**		**67%**		**33%**

Assigning meanings to the created clusters

The basic premise of assigning meanings to the created clusters is that shots within the same cluster share similar meanings In our work, the cluster meanings are represented by index terms which come from the automatic speech recognition (ASR) of words spoken in the member shots for the cluster These spoken terms are the only source of semantic information for our experiments although other sources are possible such as OCR of any printed captions or text appearing in the video A modified TF*IDF weighting schema is applied to assign each spoken term to a weight (see Formula 6-3)

A previous example cluster from video 133 (Figure 6-5) can be used to illustrate the cluster meanings (see Table 6-10 below) The meaning of cluster133_18 is represented by a great number of terms and tends to cover a variety of topics Some

persons' names were mentioned in the cluster (e g "jim", "chandler", "georg") and some business terms were seen (e g "mortgag", "banker", "commission")

Table 6-10 The meanings of example cluster 133_18

Video ID & Cluster ID	terms derived from clusters
Video133 "Road to Better Living, The"	mortgag, chandler, banker, georg, christi, jim, sound, fine, thing, bacterium, commission, plateau, telecharg, busi, castro, leonard, secretari, ken, mercer, citi, fingertip, pope, maine, freight, jame, corpor, steven, work, gentleman, depress, scholar, vice, map, elect, draw, council, charg, credit, chamber, rout, commerc, locat, loan, promis, coupl, touch, fall, moment, polic, model, factori, road, rate, game, manag, progress, inform, half, mother, scene, govern, idea, engin, commun, power, fact, compani, area, line, man, product, life

When a narrator comments on a sequence of shots, the spoken terms are a description of what we see in the sequence at that point in time, as shown in examples A problem with the process is that as the number of shots within a cluster increases the number of terms to describe the cluster also quickly increases and as a consequence important cluster meanings become less prominent

Also worth mentioning is that TRECVID2002 videos are documentation and hence informative and descriptive in nature, whereas movies and sit-coms are emotive, they do not describe in the dialogue, what is on-screen

TV news programmes are factually grounded but they are not always descriptive For instance, an anchor shot shows a news reporter on the screen and the corresponding text annotation is not about the anchor person but a summary of a piece of news

Inferring shot term weights from cluster meanings

The purpose of this process is to enhance the meaning of a shot by adding a certain level of description based on the cluster meanings (see Formula 6-4) The original ASR transcripts associated with a shot are often limited and it is hoped that the introduction of cluster meanings will help us understand more about the shot based on the assumption of clustering – that shots within the same cluster present weak semantic similarity

A major advantage of this approach is that silent shots from videos which have no ASR could receive some semantic terms from their similarity to shots in the same cluster which have an ASR associated with them. There are times when shots within some groupings are not semantically similar. Naturally shots in these groupings will not benefit from the addition of cluster terms, but would create misleading information about the shot.

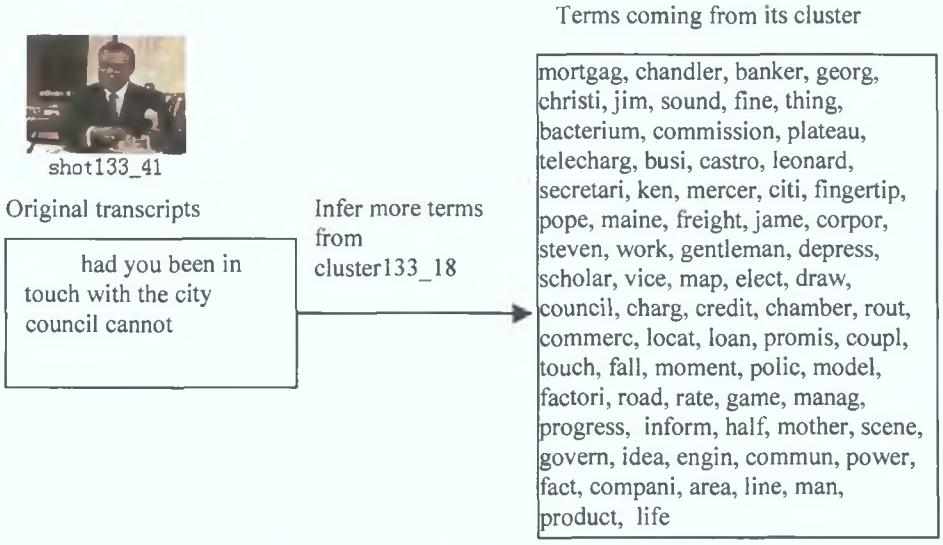


Figure 6-9: Inferring more terms from cluster meanings for shot 133_41

We take shot133_41 as an example to explain the idea (see Figure 6-9). Shot133_41 was placed into cluster133_18 in Figure 6-5. The original ASR transcript of the shot - "had you been in touch with the city council cannot", did not tell us much information of what the shot was about apart from knowing "city" and "council". What the original transcript lacks is information of what we can really see in the shot. After its cluster meanings were linked to the shot, we know more about the shot, for instance, the person who is holding a phone set may be "Chandler" or "George" and he is saying things about "mortgage" and "credit".

Mapping a non-text query onto a term vector query

Non-text queries are queries that are formulated in terms of content-based /concept-based features. They are normally represented as feature vectors. The conventional way of treating the non-text queries in video retrieval is to compare and rank the distances between a given non-text query and all searchable units in the collection.

The searchable unit can be a shot or a cluster centre if shot clustering is applied. In our work, the searchable unit is the cluster centre.

Based on the same assumption of the clustering process, given a non-text query, if the *topK* closest clusters are returned (where *topK* is the number of the closest clusters to the query), the shots within the *topK* clusters would be very similar to the query in terms of content-based/concept-based features. Following this, the meanings of the chosen *topK* clusters can be regarded as a text description of the non-text query. The text description can be used in the same way as a text query to return relevant shots based on the vector space retrieval model.

Generating a text description of a non-text query is similar to an automatic query expansion process in text retrieval. The traditional view of automatic query expansion is to expand a given text query based on the terms found in the top *k* relevant documents returned. The use of previously retrieved document texts suggests additional or alternative possible query terms. Following the similar idea, we use a non-text query for query expansion in order to find additional query terms based on the texts of the *topK* similar clusters to the query. The *topK* clusters are withdrawn after its text description is obtained and the text description continues the search.

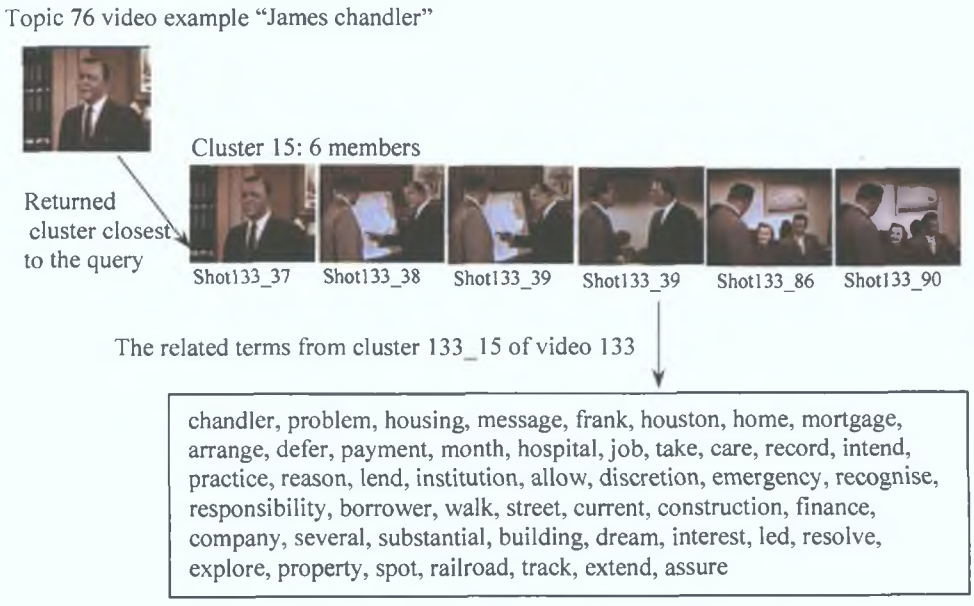


Figure 6-10: Mapping a non-text query onto a text description

The above example shows the non-text query derivation process, in particular using a content-based query To obtain the content-based features of the given video example, we first extracted its key frame and automatically calculated the required content-based features – regional dominant colour, regional colour histogram and edge histogram 3 video examples were given for Topic 76 "Find shots with James Chandler in them" Shown in Figure 6-10 is one of the video examples and a returned cluster closest to it The text description related to the returned cluster is listed in the text box It can be seen that the text description for the given query is expanded to include the terms "mortgage" and "borrower" Also found is the term "Chandler"

6.4 Experiments

Video consists of three categories of features spoken-text, concept-based and content-based features In addition to being the primary sources for video retrieval these features are the main complication of creating and manipulating a searchable index An outcome of the complication is that many combinations of features are possible and analysis is required in order to decide which features to use for video retrieval and how to use them Our experiments will concentrate on the 9 systems listed in Table 6-7

Section 6 4 1 will examine four systems and discuss the performances of using a combined index built from concept-based and content-based features in conjunction with spoken-text feature Whether a combined query inferred from a non-text query along with the original text query is feasible will be illustrated for each of the 3 systems in section 6 4 2 Also discussed is a comparison of the performance difference between using one best image example and using all image examples in a combined query The comparison will take Sys4 as an example to consider content-based features only in section 6 4 3 Finally, back to the analysis of non spoken-text features in video retrieval, systems Sys3 and Sys6 will look into the performance of using content-based and concept-based alone queries through a combined index, respectively

6.4.1 Retrieval Performance When Using an Aggregated Index

The analysis of results given by four systems helps us understand the effects of using a combined index Sys1, Sys2, Sys5 and Sys8 Only the original text queries were used for the comparison (see Table 6-5) The combined index is generated differently as shown below

- Sys1 our baseline system using a traditional TF*IDF index,
- Sys2 an aggregated index from conventional TF*IDF term weights and derived term weights from content-based features,
- Sys5 an aggregated index from conventional TF*IDF term weights and derived term weights from concept-based features,
- Sys8 an aggregated index from conventional TF*IDF term weights and derived term weights from content-based features and concept-based features

Before presenting an analysis from amongst the four systems, we first compare the performances of Sys2 and Sys5 produced when varying just two parameters We then take the best clustering results according to performances in order to complete the combined index for Sys8

Two variables under consideration are the temporal window size *WIN* and threshold xT as used in the clustering process The chosen 10 value pairs for variable *WIN* and xT were listed earlier in Table 6-7 There were 10 runs for Sys2 and Sys5, respectively The random selection of cluster centres made the resulting clusters for each run distinct

Table 6-11 below lists values for Mean Average Precision when varying *WIN* and xT for Sys2 and Sys5 Mean Average Precision (MAP) is the mean value of Average Precision over the 25 TRECVID2002 topics and it measures the overall performance among the runs for each system Higher MAP values indicates better system performance

Table 6-11 Mean average precision by variable WIN and xT for Sys2 and Sys5

System	WIN =0	WIN = 1			WIN = 2			WIN =3		
		xT=0 1	xT= 0 15	xT=0 2	xT=0 1	xT= 0 15	xT=0 2	xT=0 1	xT= 0 15	xT=0 2
Sys2	0 075	0 072	0 076	0 076	**0 078**	0 076	0 069	0 067	0 068	0 074
Sys5	0 080	0 078	0 078	0 064	**0 078**	0 073	0 073	0 073	0 071	0 082

Sys2

The MAP has a maximum value at 0 078 for *WIN* = 2 and it then declines gradually after *WIN* = 2 *WIN* and *xT* account for the thresholds that are needed to determine the dissimilarity between any given shot and their neighbouring shots The reason of using the variable pair is to put the shot and its very similar neighbours into the same cluster so as to create a cluster containing shots with visual similarity However, the introduction of *WIN* and *xT* did not give any performance benefit in Sys2 since the MAP differences are so small

Sys5

The MAP value stays above 0 07 for most variable pairs No obvious performance difference is found in Sys5 yet peak MAP was reached at 0 082 when *WIN*=3 and *xT*=0 2 No definite decisions about which pair of *WIN* and *xT* worked best can be made from the experiments and this suggests that results obtained from temporal shot clustering are very comparable to results from the non-temporal shot clustering in Sys5

It is also seen that the performance of Sys5 is comparable to Sys2 over most variable pairs No conclusion can be drawn whether clustering by using concept-based features groups shots slightly better than using content-based features

Sys1, Sys2, Sys5 vs Sys8

We took the best results produced by Sys2 and Sys5 in order to complete the combined index needed for Sys8 The resulting clusters generated were based on the following variable setting *WIN*=2 and *xT*=0 1 Table 6-12 below illustrates a comparison of the overall performance among the four systems Sys1, Sys2, Sys5 and Sys8

It can be seen that MAP of Sys2, Sys5 and Sys8 are comparable to each other, and higher than the baseline system Sys1 as anticipated from the earlier discussion. Sys2, Sys5 and Sys8 retrieved more relevant shots over the 25 topics. It shows that an aggregated index derived from the spoken-text feature and content-based /concept-based features provides marginal improvements in a video retrieval system.

Table 6-12: Summary of performance of Sys1, Sys2, Sys5 and Sys8

System	MAP	Total relevant shots retrieved
Sys1	0.056	139
Sys2	0.078	176
Sys5	0.078	164
Sys8	0.076	166

Figure 6-11 below shows the precision at 11-recall levels for the four systems. At 0 recall level, Sys1 has higher MAP than the other systems. After that, Sys2, Sys5 and Sys8 take the lead over Sys1 all the way through to recall at 1.0. These three systems are comparable to each other. Figure 6-12 shows the precision at 6 document cut-off levels. All systems perform similarly at the 5 document cut-off level and after that the precision of Sys1 starts sliding down more rapidly than the other systems.

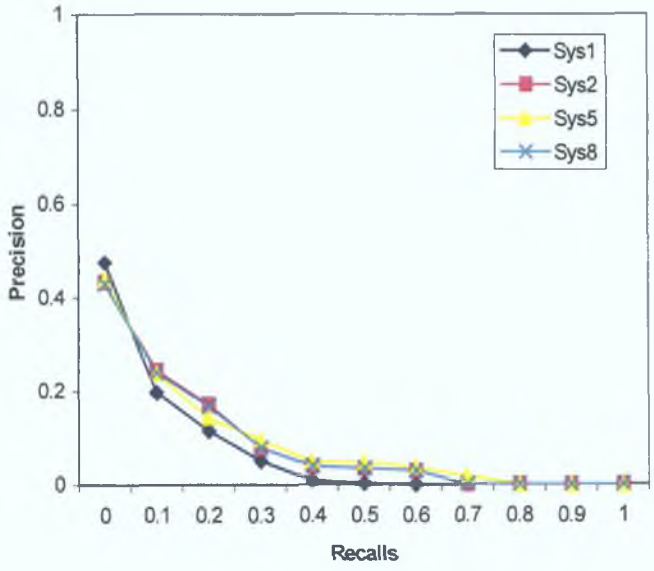


Figure 6-11: Precision at recalls for Sys1, Sys2, Sys5 and Sys8

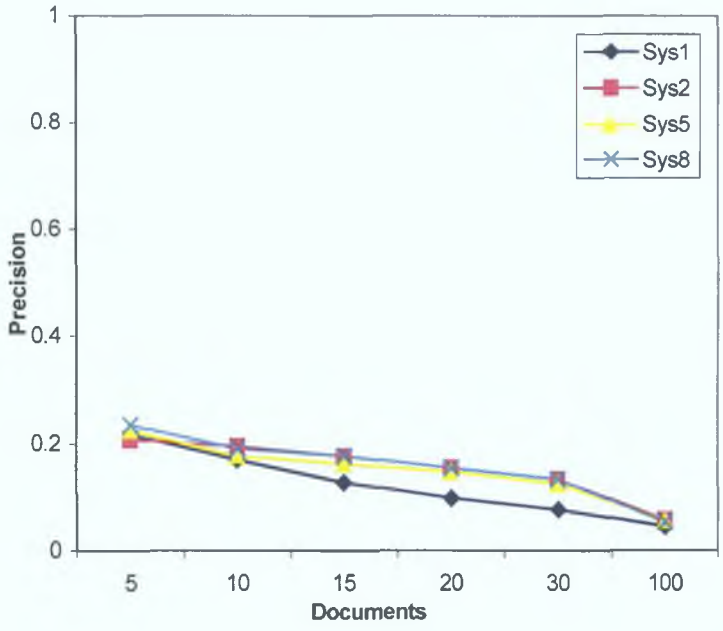


Figure 6-12: Precision at document cut-offs for Sys1, Sys2, Sys5 and Sys8

These results show that an aggregated indexing method yields marginally improved results over the traditional TF*IDF indexing method based on ASR for video retrieval. The improvement in recall leads to a slight reduction of precision at the 0 recall level. In other words, the aggregated index does retrieve more relevant shots but could not improve their ranking. It is shown that the introduction of cluster meanings to its member shots not only enhances the shot meanings but also weakens them to some extent.

The result of Sys8 indicates that a retrieval system using a combined index with all three different types of features did not outperform one applying a combined index with two types of features: spoken-text and content-based features (or spoken-text and concept-based features). The derived meanings of a shot in Sys8 come from two different clusters: one created based on content-based features and the other based on concept-based features. If these two clusters are semantically similar enough, the meanings of the shot should be strengthened. But the fact is that the two clusters are created based on different features and their cluster semantics may be different and it would instead weaken the meanings of the shot.

Furthermore, the performance graphs for Sys2 and Sys8 are very similar and the clustering results of Sys2 plays an important part in Sys8 indicating that the clustering results of Sys2 are better than that of Sys5 in terms of semantic grouping regardless the features used for grouping.

Figure 6-13 below illustrates the average precision of each TRECVID2002 topics for four systems and the median of average precision (noted as TREC_median) by TRECVID2002 participants. It can be seen that, except for topics 76 and 84, all systems worked very comparably for most topics. Sys2 and Sys5 provide essentially the highest average precision for topic 76 on "James Chandler" and second highest for topic 84 on "Price Tower".

The good performance of these topics indicates that the process of inferring additional meaning for a shot from its associated cluster can facilitate useful indexing. To take an example of "James Chandler", a shot could be showing James talking with people in his office but no "James" or "Chandlers" is indexed for the shot. The shot was given term "James" and/or "Chandler" after it was clustered and associated with other similar shots which contain these terms.

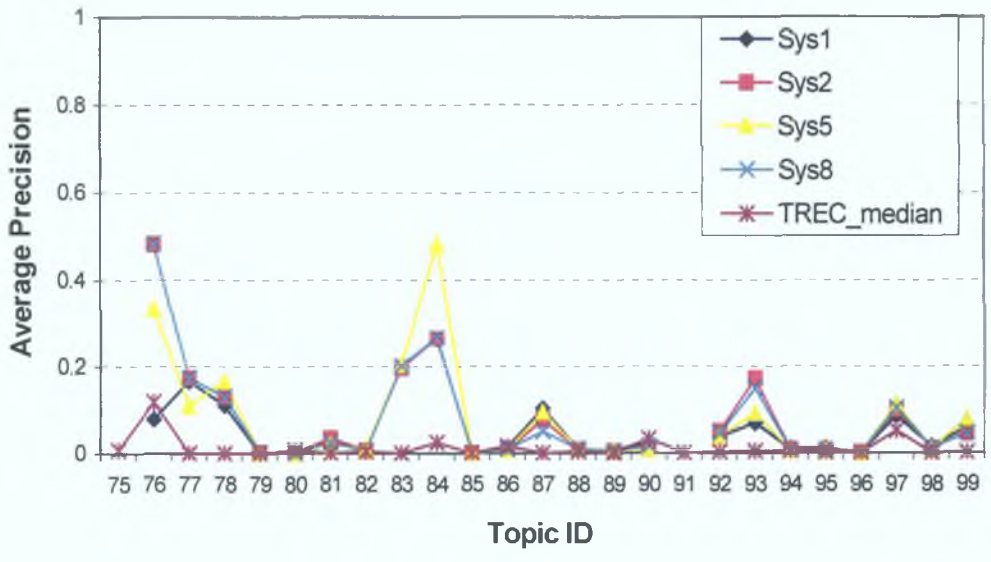


Figure 6-13: Average precision per topic for Sys1, Sys2, Sys5, Sys8 and TREC_median

6.4 2 Retrieval Performance When Using an Aggregated Query

This section continues discussing the feasibility of using an aggregated query in video retrieval The aggregated query is first derived from a concept-based (or content-based) query and then aggregated with the original text query The systems used in this comparison are Sys2, Sys4, Sys5, Sys7, Sys8 and Sys9 We paired up Sys2 and Sys4, Sys5 and Sys7, and finally Sys8 and Sys9 for study, respectively

The study of variables *WIN* and *xT* told us that a temporal clustering algorithm and a non-temporal one performs similarly in section 6 4 1 We chose one set of clustering results created when *WIN*= 2 and *xT* = 0 1 for content-based and concept-based features, respectively

Two variables are under consideration for Sys4, Sys7 and Sys9 (1) the number of most similar clusters chosen to map a content-based query onto a query term vector - *topK*, (2) the weight *PF* given to an original text query when creating a combined query The number of *topK* is chosen to be 1, 3, 5, 7 or 9, and the weight *PF* is given to be 4, 8, 12, 16, 20 or 24

Sys2 vs Sys4

Sys2 is treated as the baseline system for Sys4 Both systems use the same aggregated index It was constructed in such a way as to combine conventional TF*IDF term weights of shots and the term weights derived from content-based features The only difference is that Sys2 uses text-only queries while Sys4 uses an aggregated query that comes from the text and content-based query Content-based queries are in a term vector form which is obtained based on content-based features of one manually picked best image example

To compare Sys2 (Table 6-11) and Sys4 (Table 6-13 below), the maximum MAP for Sys2 is 0 078 and 0 084 for Sys4 It implies that using an aggregated query for video retrieval gives slight performance benefit over using a text-only query The improvement of Sys4 over Sys2 contributes to the introduction of the non-text query derivation process (see later section 6 4 4) is achieved by stressing the importance of the original text query (where weight *PF* was 20) in order to improve

precision, meanwhile by keeping all extra query terms derived from the non-text query to low weight in order to improve recall

Table 6-13 below shows that there is no performance benefit in introducing the variable *topK* The performance remains almost the same as the value for *topK* increases for each given *PF* Variable *topK* is related to the number of most similar clusters to a given content-based query The results indicate the possibility of finding the *topK* most similar clusters using content-based feature measures This is because content-based representations of a query are specific enough to distinguish relevant shots from the collection A problem still remains however if no correct clusters were found, the derived query could be poorly formed where the derived query terms could be unrelated to a given topic

Table 6-13 Mean average precision by variable topK and PF for Sys4 using one best image example in a query

TopK	Weight PF					
	4	8	12	16	20	24
1	0 0658	0 0702	0 0714	0 0723	0 0730	0 0734
3	0 0638	0 0723	0 0766	0 0788	0 0794	0 0796
5	0 0667	0 0763	0 0804	0 0825	0 0833	0 0835
7	0 0640	0 0764	0 0816	0 0834	0 0842	**0 0846**
9	0 0636	0 0754	0 0786	0 0816	0 0824	0 0827

An increase was found in MAP as parameter *PF* increases but slows down after *PF*= 16 Variable *PF* defines the proportion of weight given to an original text query when creating a combined query The process of deriving term vectors for non-text queries introduces more unrelated terms to a given topic A higher value is assigned to the original query terms than the derived terms before combining them Variable *PF* helps maintain the importance of original text query terms and include additional terms derived from non-text queries

Sys5 vs Sys7

Sys5 is treated as the baseline system for Sys7 Both systems use the same aggregated index that is a combination of conventional TF*IDF term weights and the term weights derived from concept-based features The distinction is that Sys5 uses text-only queries while Sys7 uses an aggregated query that comes from a text and concept-based query The concept-based query is in a term vector form derived based on the concept-based queries (see Table 6-5)

Comparing the chosen MAP 0 078 of Sys5 (Table 6-11) and the best 0 076 of Sys7 (Table 6-14 below), it is shown that using an aggregated query that comes from a concept-based query and a text-only query for video retrieval can maintain the same performance level as using a text-only query but can not go beyond This is due to the generality of the concepts defined by TRECVID2002 where the concept queries were too general to find any appropriate or relevant *topK* similar clusters for query derivation as shown in later section 6 4 4 Even though the derived query terms are not useful in expressing the topic the performance of Sys7 was maintained by assigning more important values to original query terms prior to the query aggregation

Table 6-14 Mean average precision by variable topK and PF for Sys7

TopK	Weight PF					
	4	8	12	16	20	24
1	0 0676	0 0746	0 0760	0 0762	**0 0764**	0 0764
3	0 0637	0 0727	0 0751	0 0761	0 0763	0 0763
5	0 0627	0 0728	0 0746	0 0752	0 0759	0 0762
7	0 0635	0 0726	0 0742	0 0753	0 0754	0 0755
9	0 0605	0 0688	0 0734	0 0740	0 0752	0 0753

Sys8 vs Sys9

Sys8 and Sys9 used an aggregated index from conventional TF*IDF term weights and the derived term weights from content-based features and concept-based features Sys9 was created by taking the resulting clusters generated based on variable *WIN*=2 and *xT*=0 1 from Sys2 and Sys5 The aggregated queries for Sys9 was a combination of text-only, concept-based and concept-based query

Table 6-15 Mean average precision by variable topK and PF for Sys9

TopK	Weight PF					
	4	8	12	16	20	24
1	0 0573	0 0573	0 0573	0 0573	0 0573	0 0573
3	0 0574	0 0681	0 0723	0 0733	0 0738	0 0728
5	0 0567	0 0695	0 0723	0 0735	0 0742	**0 0744**
7	0 0533	0 0685	0 0701	0 0717	0 0722	0 0724
9	0 0514	0 0663	0 0704	0 0712	0 0719	0 0721

The MAP for Sys8 is 0 076 and 0 074 the best for Sys9 (see Table 6-15 above) The result shows that using an aggregated query that comes from text-only, concept-based and concept-based query for video retrieval can also maintain the same performance level as using a text-only query but can not go beyond

Sys2, Sys4, Sys5, Sys7, Sys8 vs Sys9

We took the best result with the highest MAP of each system for comparison Sys2, Sys4, Sys7, Sys8 and Sys9 Table 6-16 below shows the MAP among the systems and the total relevant shots retrieved over the 25 topics along with their specific baseline systems Figure 6-14 below plots the precision at 11 recall levels Figure 6-15 below gives the precision at 6 document cut-off levels

Table 6-16 Summary of the best performance of the six systems

	Sys1	**Sys2**	**Sys4**	**Sys5**	**Sys7**	**Sys8**	**Sys9**
MAP	0 056	0 078	0 084	0 078	0 076	0 076	0 074
Total relevant shots retrieved	139	176	186	164	171	166	174

Results show that all systems are very comparable indicating that the usage of a combined query derived from a non-text query is possible Sys4, Sys7 and Sys9 retrieved more relevant shots than their corresponding baseline systems Sys2, Sys5 and Sys8 The query derivation process introduces a number of query terms unrelated to topics The original text terms play a major role in video retrieval and given most weights in order to focus the meaning of a given topic

The small loss in MAP for Sys7 and Sys9 was caused by the difficulty of finding the correct clusters for query derivation using concept-based features The generality of the defined concepts does not make it easy to construct a query that has the ability to distinguish clusters

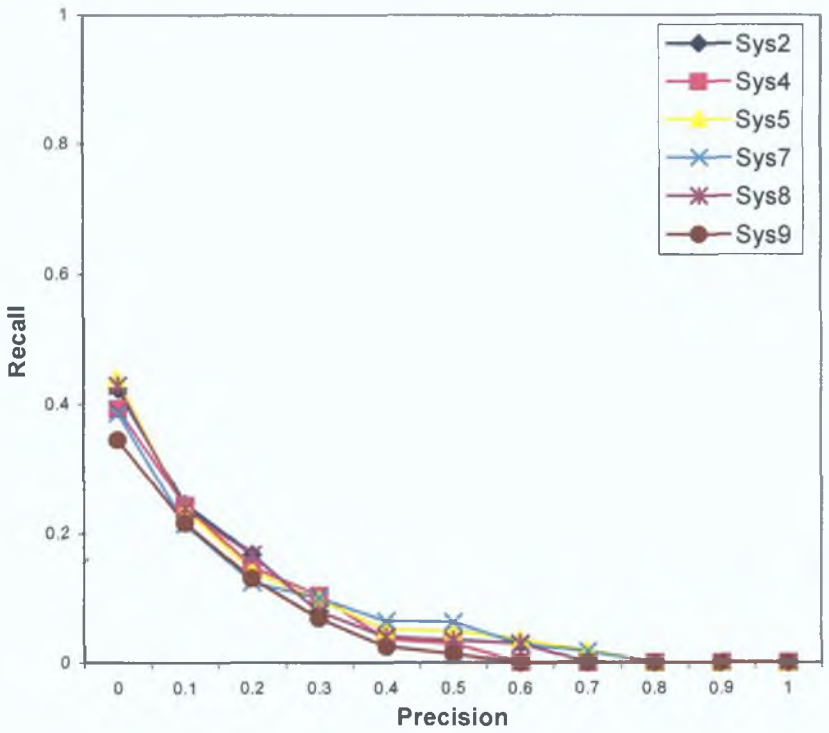


Figure 6-14: Precision at recalls for 6 systems

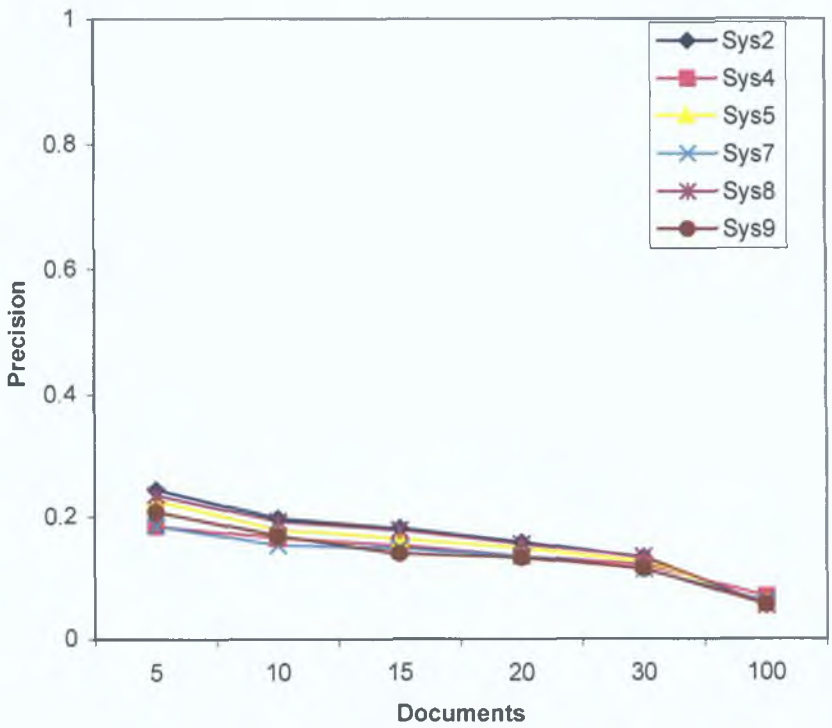


Figure 6-15: Precision at document cut-offs for 6 systems

Figure 6-16 shows the average precision per topic over the 6 systems. Sys4 did well in topic 75 "Eddie Rickenbacker" and topic 76 "James Chandler", both of which have image examples selected from the same collection. Choosing images from a comparable collection as a content-based query is preferable.

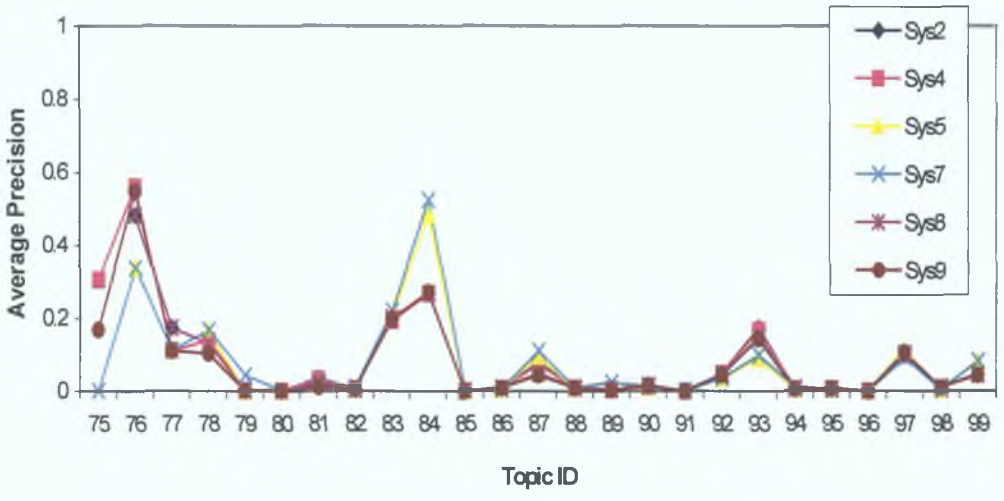


Figure 6-16: Average precision by topic for 6 systems

6.4.3 Retrieval Performance When Using Multiple Image Examples in a Query

The purpose of this section is to examine the performance difference between using one best image example and using all image examples in a query. Only queries using content-based features are under consideration in this experiment where derived text description combined with the original query terms are applied to generate the final query to the systems. The thinking here is to tackle the problem of manually picking one best image query for video retrieval which is difficult if users have no knowledge about the search collection, namely the quality of pictures, or the retrieval systems. We used two types of query formulation in a query for Sys4:

(1) Sys4_1: one image example

The image example was manually chosen based on the assumption that we have certain background information about the search collection. The selection criterion is that the chosen image should be very comparable to the search collection in terms of picture quality. If no such image example were given by

TRECVID2002 for a topic we randomly picked it. For instance, topic 92 of "sailboats or clipper ships" was given six examples (Figure 6-17 below): four are images from outside the collection and the others are video clips from within the TRECVID2002 test collection, and we chose one of the video clips as the best image example in the query for the topic.

(a) images from outside the collection

(b) images from inside the collection

Figure 6-17: Six visual examples given for topic 92 of "sailboats or clipper ships"

(2) Sys4_2: multiple image examples

In this variation, all image / video examples provided by TRECVID2002 are used in the query for a given topic. Formula 6-7 showed how to combine the multiple images into a single text description.

Table 6-17 below shows the performance of Sys4_2 when varying parameters *topK* and *PF*. It is found that there is no significant performance difference for variable *topK* and there is a tendency for MAP to increase as the *PF* weight increases.

Table 6-17 Mean average precision by variable topK and PF for Sys4_2 using multiple image examples in a query

TopK	Weight PF					
	4	8	12	16	20	24
1	0 0589	0 0668	0 0683	0 0694	0 0713	0 0740
3	0 0569	0 0710	0 0748	0 0765	0 0774	0 0775
5	0 0558	0 0716	0 0752	0 0763	0 0774	**0 0788**
7	0 0511	0 0708	0 0757	0 0758	0 0766	0 0771
9	0 0486	0 0642	0 0721	0 0746	0 0742	0 0744

To compare Table 6-13 in section 6 4 2 and Table 6 17, Sys4_1 performs better than Sys4_2 over all variable settings suggesting that using multiple images in a query for video retrieval introduces more unrelated query terms if no correct clusters are found for each image example The additional query terms appear to cause a slight reduction in precision

Comparing the best results for Sys4_1 in Table 6-13 with the best results from Sys4_2 in Table 6-18 below, both systems have similar performance The more image examples are used in a query, the more query terms derived are unrelated to a given topic It is crucial to maintain the importance of original text query terms while additional query terms from non-text queries are added in

Table 6-18 Summary of the best performance of Sys4_1 and Sy4_2

	Sys4_1	Sys4_2
MAP	0 084	0 079
Total relevant shots retrieved	186	183

Figures 6-18 and 6-19 below plot the precision at 11 recall levels and at 6 document cut-off levels for each system, respectively Figure 6-20 gives the average precision by topic for Sys4_1 and Sys4_2 The graph shows the two systems are similar in performance

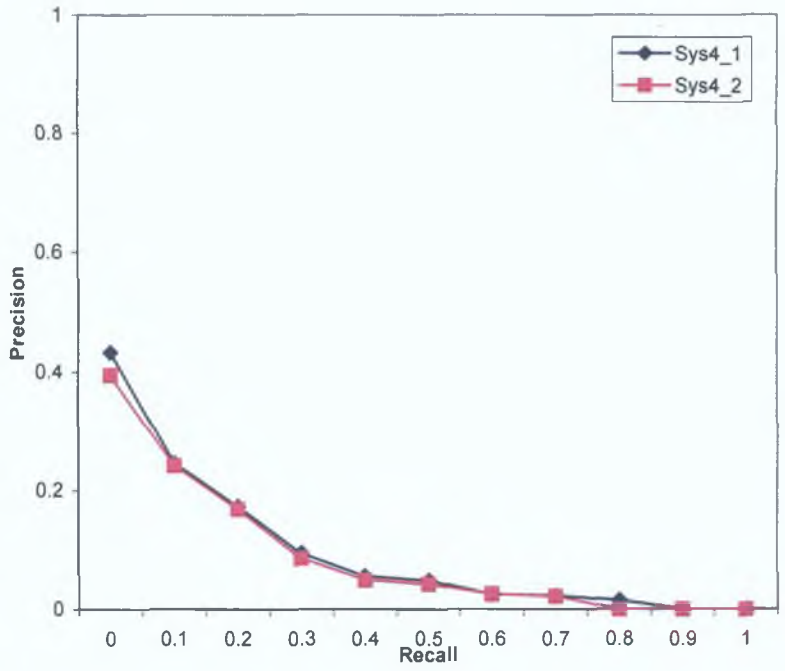


Figure 6-18: Precision at recalls for Sys4_1 and Sys4_2

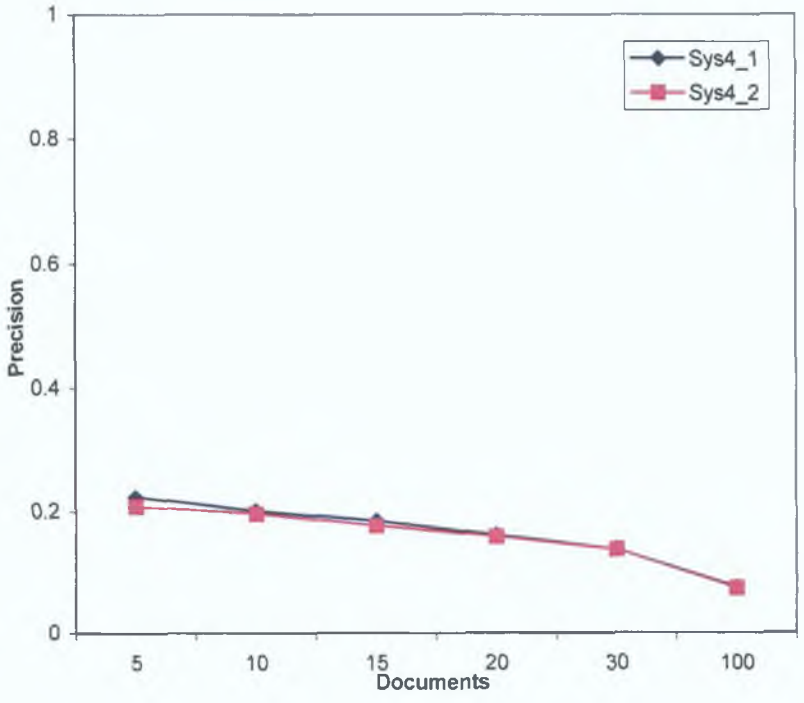


Figure 6-19: Precision at document cut-offs for Sys4_1 and Sys4_2

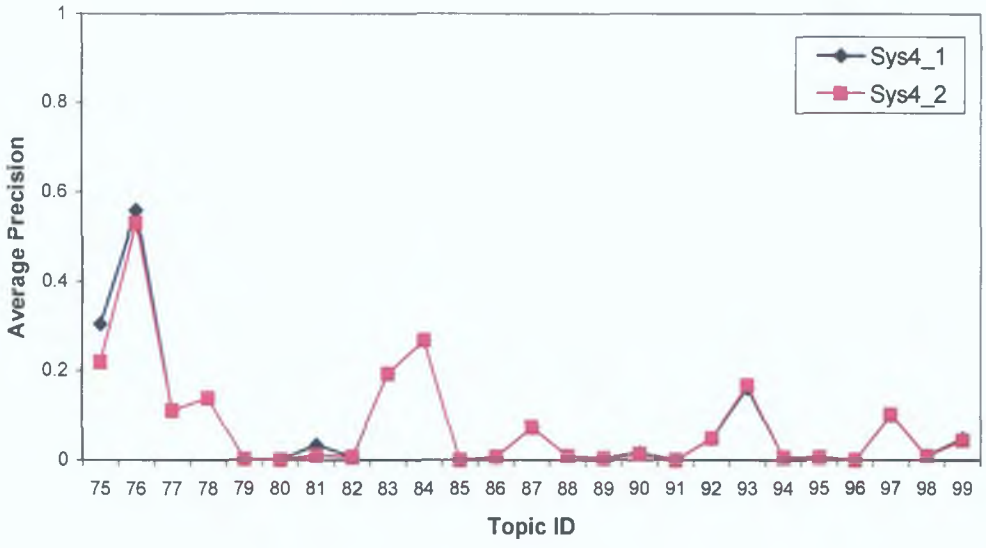


Figure 6-20: Average precision per topic for Sys4_1 and Sys4_2

6.4.4 Retrieval Performance When Using Non-Spoken Text Features

Sys3 and Sys6 are additional video retrieval systems that show the performance difference when using a content-based only and concept-based only query for video retrieval. Each system is set up as follows.

Sys3 uses a combined index that combines conventional TF*IDF term weights of shots and the term weights derived from content-based features. A combined query is derived based on content-based query. As mentioned in section 6.4.3, there are two different ways to use the image examples for formulating a content-based query. We therefore examine Sys3 with two different query formulation settings, one using one manually-chosen image example in a query is noted as Sys3_1 and the other one using multiple image examples where provided in TRECVID2002 topic descriptions is noted as Sys3_2.

It can be seen that Sys3_1 performs slightly better than Sys3_2 (see Table 6-19 and 6-20 below) suggesting that using one best image in a query for video retrieval is preferable. Using multiple images in a query is possible when users have no knowledge about the search collection.

Table 6-19 Mean average precision by variable topK, WIN and xT for Sys3_1 using one image example in a query

TopK	WIN =0	WIN = 1			WIN = 2			WIN =3		
		xT=0 1	xT= 0 15	xT=0 2	xT=0 1	xT= 0 15	xT=0 2	xT=0 1	xT= 0 15	xT=0 2
1	0 0010	0 0107	0 0001	0 0159	0 0184	0 0114	0 0002	0 0202	0 0168	0 0003
3	0 0004	0 0073	0 0048	0 0087	0 0251	0 0102	0 0038	0 0170	0 0177	0 0223
5	0 0018	0 0150	0 0049	0 0075	**0 0282**	0 0097	0 0034	0 0206	0 0174	0 0165
7	0 0005	0 0144	0 0106	0 0062	0 0276	0 0024	0 0142	0 0213	0 0167	0 0138
9	0 0004	0 0146	0 0130	0 0052	0 0257	0 0018	0 0126	0 0212	0 0166	0 0079

Table 6-20 Mean average precision by variable topK, WIN and xT for Sys3_2 using multiple image examples in a query

TopK	WIN =0	WIN = 1			WIN = 2			WIN =3		
		xT=0 1	xT= 0 15	xT=0 2	xT=0 1	xT= 0 15	xT=0 2	xT=0 1	xT= 0 15	xT=0 2
1	0 0003	0 0092	0 0020	0 0085	0 0165	0 0039	0 0002	0 0195	0 0092	0 0113
3	0 0006	0 0073	0 0035	0 0077	**0 0210**	0 0063	0 0020	0 0166	0 0086	0 0194
5	0 0058	0 0132	0 0033	0 0075	0 0201	0 0037	0 0015	0 0136	0 0079	0 0148
7	0 0060	0 0128	0 0090	0 0039	0 0208	0 0027	0 0054	0 0141	0 0095	0 0089
9	0 0060	0 0128	0 0078	0 0037	0 0204	0 0025	0 0049	0 0135	0 0087	0 0052

Table 6-21 Mean average precision by variable topK, WIN and xT for Sys6 using one image example in a query

TopK	WIN =0	WIN = 1			WIN = 2			WIN =3		
		xT=0 1	xT= 0 15	xT=0 2	xT=0 1	xT= 0 15	xT=0 2	xT=0 1	xT= 0 15	xT=0 2
1	0 0004	0	0 0005	0 0014	0 0001	0 0009	0 0014	0 0022	0 0001	0 0012
3	0 0018	0 0002	0 0014	0 0072	0 0050	0 0007	0 0030	0 0008	0 0008	0 0010
5	0 0017	0 0003	0 0019	0 0027	0 0062	0 0004	0 0026	0 0014	0 0007	0 0012
7	0 0026	0 0004	0 0029	0 0018	0 0083	0 0005	0 0022	0 0016	0 0005	0 0012
9	0 0023	0 0005	0 0040	0 0016	**0 0088**	0 0005	0 0029	0 0019	0 0004	0 0023

Sys6 uses an aggregated index that combines conventional TF*IDF term weights of shots and the term weights derived from concept-based features An aggregated query is derived based on a concept-based query The result in Table 6-21 above shows that the performance of Sys6 is very poor - the best MAP is only 0 0088 Using a derived text query from a concept-based query in video retrieval is likely to be poor because of the difficulty of finding the *topK* most similar clusters using concept-based features Here, the concepts are so general that clusters can not be distinguished Furthermore, the construction of concept-based query using 3 scales {0, 0 5, 1} allows the probability of formulating the same textual query for different topics

Sys3_1, Sys3_2 vs. Sys6

Table 6-22 below summarises the performance of the best runs for Sys3_1, Sys3_2 and Sys6. The Mean Average Precision (MAP) for Sys3_1 and Sys_2 are higher than for Sys6. They also retrieved more relevant shots than Sys6. Figures 6-21 and 6-22 below plot the precision at 11 recall levels and at 6 document cut-off levels. It can be seen that Sys3_1 and Sys3_2 have better performance than Sys6, suggesting that using a derived text query from content-based query is feasible in video retrieval but it is not necessary for systems to use a derived text query from concept-base query in video retrieval.

Table 6-22: Summary of the performance of Sys3_1, Sys3_2 and Sy6

	Sys3_1	Sys3_2	Sys6
MAP	0.0282	0.0210	0.0088
Total relevant shots retrieved	47	36	27

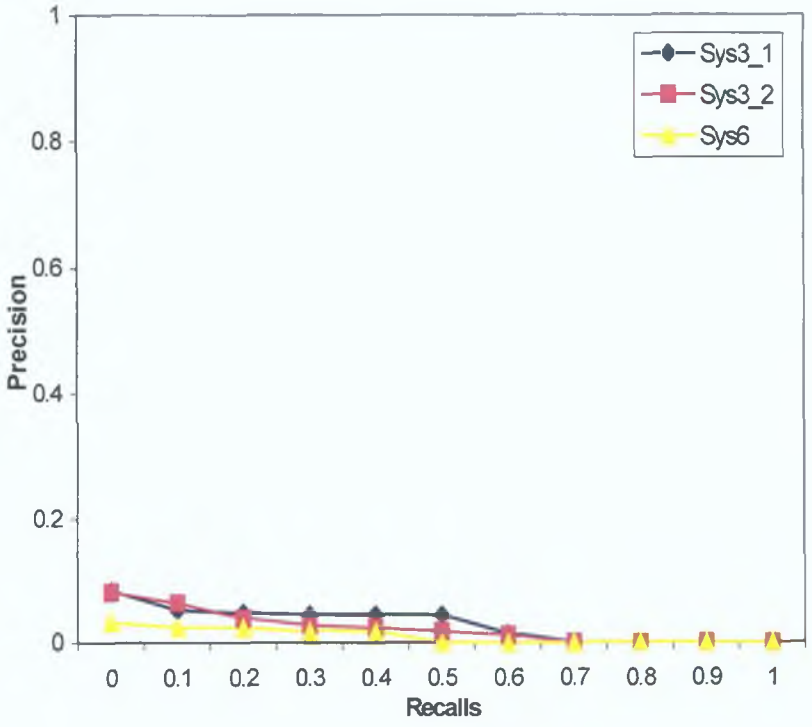


Figure 6-21: Precision at recalls for Sys3_1, Sys3_2 and Sys6

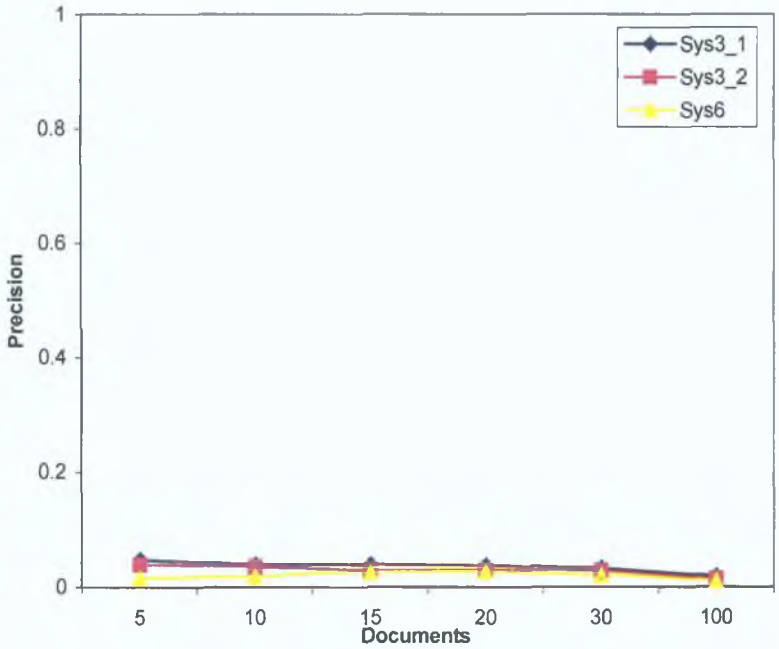


Figure 6-22: Precision at document cut-offs for Sys3-1, Sys3_2 and Sys6

Three systems retrieved a small amount of relevant shots for a very limited number of topics. Sys3_1 and Sys3_2 perform well for Topics 75 "Eddie Rickenbacker" and 76 "James Chandler" (see Figure 6-23 below). This is because the image query examples given for both of these topics by TRECVID2002 are from within the same search collection making it easier to find results than topic examples outside the collection. The derived text queries for both topics contain terms relevant to the information need.

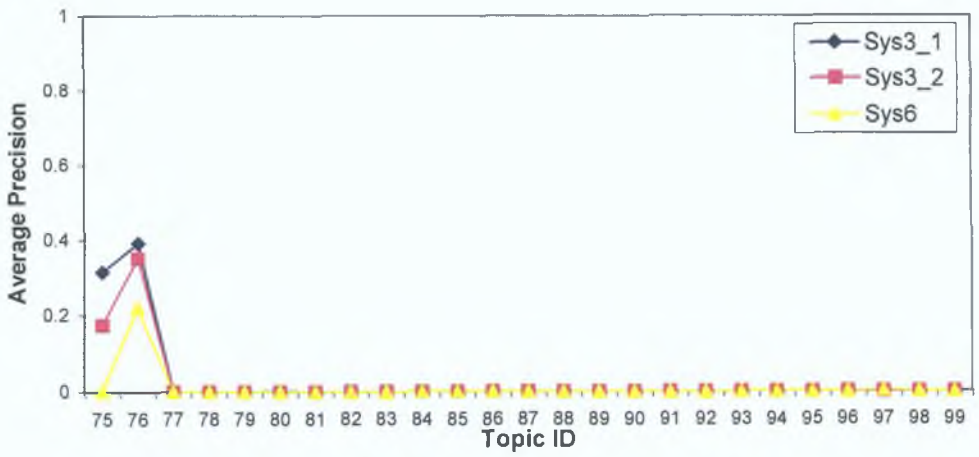


Figure 6-23: Average precision per topic for Sys3_1, Sys3_2 and Sys6

6.5 Conclusions

MPEG-7 is a generic standard used to encode information about multimedia content and often, different MPEG-7 Descriptor Schemas are instantiated for different representations of a shot such as text annotations and visual features Our work in this thesis focuses on two main areas, the first is devising a method for combining text annotations and visual features into one single MPEG-7 description and the second is how best to carry out text and non-text queries for retrieval via a combined description

One of the challenges with video retrieval results from the difficulty of combining different types of features that can be automatically detected from videos and their detection accuracy In the work we have reported here, we define and classify the features (which are equivalent to MPEG-7 descriptions) into three main categories (1) spoken-text features, (2) concept-based features, (3) content-based features Although these features can be computed and represented by symbols (i e characters or vectors) it is these representations that determine how well they can approximate the semantic world

Of the three types of features, spoken-text features are the most precise representation of shot content due to the rich semantic information they contain The frequency of term occurrence in a shot is a useful measurement for determining the meaning of the shot

Content-based features are simply vector representations of shots and limited semantics can be directly obtained from these vectors The poor performance of using content-based only features in our experiments for video retrieval is due to a lack of detailed and accurate semantic information in those features Concept-based features are therefore introduced by replacing content-based vectors with concepts in order to try to bridge the gap between semantics and content-based features However the number of semantic concepts is so large that to construct a reasonable set of concept-based features would require a great amount of effort in designing the concepts and the mapping models

A possible way of using content-based or concept-based features in video retrieval is to modify the video retrieval processing to a text retrieval process, where a semantic matching and measuring technique is already in place, possibly a vector space model We assume that shots in the same cluster are not only similar visually but also semantically to a certain extent Our attempt is to map non-text features onto a text description via a clustering process The text description is a term weight vector which can then be combined with spoken-text features to produce a final searchable index

Our TRECVID2002 experiments on using an aggregated index from spoken-text and content-based/ concept-based features for video retrieval validate the assumption of the visual and semantic similarity of shots in the same cluster The weak semantic similarity within a cluster can be used to enrich the original meanings of its constituent shots The meaning enrichment of a shot is completed by introducing other terms from its member shots An aggregated index using three types of features works in the same way as those applying only two types of features It is understood that different clustering processes create different sets of groupings and some sets preserve better meanings for the groupings than the others

This same assumption was also validated in the query formulation process showing that mapping a content-based query onto a text description and combining it with the original text query into video retrieval, is feasible But the mapping process is problematic partly due to the difficulty of finding the most similar clusters to a given query If the correct clusters to a given non-text query can be found, the mapped text description can perform in the same way as a text query If no correct clusters can be found, the derived query term vector could be poorly formed The query mapping process allows a non-text query to expand into a good description closely related to the topic or into a poor description completely different from the original topic The quality of images also affects the finding of the right clusters If a given image query is from within the clusters, an aggregated query will give performance benefit

Furthermore, it is shown that the mapping of a concept-based query onto a text description did not help video retrieval due to the generality of the defined concepts and query duplications for different topics

Another difficulty with the query mapping process is the potential of introducing more query terms that are related and unrelated to a given topic The related terms help find more relevant shots and the unrelated terms degrade the system performance This was validated by the comparison between a system using multiple images in a query and one using only one manually chosen image In order to reduce the influence of derived query terms which are unrelated to the topic, more weight is assigned to the original query terms prior to the query aggregation process The major meanings of the final aggregated query will thus remain so that its system performance can be preserved as good as one using the original query terms only

The temporal characteristic of videos allows us to add meaning to a shot based on the shots that are around it Such a way of enhancing the meaning of a shot is much easier to implement than our proposed method in which a clustering algorithm is required What our method attempts to solve is to include the semantic meaning (i e terms) from shots that are similar visually within a video to a shot These shots are not necessarily neighbours to the shot They could be from either the beginning of the video, the middle or the end Typically, a documentary film sets forth on a topic or an argument with careful and elaborate detail The shots are added as evidence to help lay out or confirm the argument They are like illustrations in a book and a great number of materials that are similar visually and semantically are accommodated throughout a documentary video

In the next Chapter we will report on re-running most of the experiments reported in Chapter 6, but using the video data collection, topics and relevance assessments from TRECVID held in 2003 This dataset, because it is significantly different from that used in TRECVID2002, will offer a contrast to the results reported in this Chapter

Chapter Seven

Experiments on the TRECVID2003 Search Collection

Our experiments on TRECVID2002 reported in Chapter 6 have shown that it is possible to enrich the semantic description of each shot in a video collection from clusters based on concept/content–based features and automatic speech recognition texts The enriched description can be used in the indexing process to improve retrieval There is little performance difference between an index derived from concept-based features and one from content-based features but in addition our use of a cluster description of semantics in the query formulation process, may be limited A query description for a topic can be derived by comparing content-based features between an image example and clusters and the derivation can be done correctly only when the given image examples are chosen from a very comparable collection A derived query description based on concept-based features did not give any performance benefit due to the difficulty of constructing a concept-based query

This chapter reports on our further experiments on the manual search task in TRECVID2003 The main distinction between TRECVID2002 and TRECVID2003 collections is that the former consists of 1930s to 1960s documentaries while the latter is comprised of 1998 broadcast TV news from CNN and ABC This chapter examines if our assumption that the semantic similarity of each shot can be determined from clusters, is still valid for collections with a different genre and image quality

Similar to the structure used in chapter 6, we will deal with 4 subtopics in our experiments here (1) an overview of the manual search task of TRECVID2003, (2) a review of the TRECVID2003 experimental results of other research groups, (3) our experimental settings and (4) our own experimental results in detail

7.1 The TRECVID 2003 Manual Search Task

The TRECVID2003 manual search task extends the previous year's in five aspects (1) the search collection, (2) the topics, (3) defined retrieval unit, (4) indexing features available for sharing, (5) manual search submission requirements

TRECVID2003 search test collection

The search data for TRECVID2003 contains 113 MPEG-1 videos, 133 hours in total In particular, the 107 videos are CNN and ABC broadcast news from April through June 1998 with commercials, weather and sports 6 videos are C-SPAN programmes from July 2001 with records of various government committee meetings, discussions of public affairs, some lectures and news conference

Table 7-1 Summary of the TRECVID2003 search collection

	ABC World News	CNN Headline News	C-SPAN
Total videos	53	54	6
Total shots	16593	15282	443

TRECVID2003 topics

There are 25 topics defined m TRECVID2003 Each topic came with a textual description of the information need along with one or more video clips or still images for illustration (see the overview Table 7-2 below) The 25 topics are about people, things, events, locations and their combinations The topics were designed by NIST to make sure that there should be multiple relevant shots for all topics and that the relevant shots come from more than one video Multimedia examples for

most topics were chosen from the same search collection or a comparable collection A few examples were from outside the collection

Table 7-2 Overview of the 25 topics of the TRECVID2003 search task

Topic ID	Textual Description of topics	Number of examples		Total Number of relevant shots
		Image	Video	
100	shots with aerial views containing both one or more buildings and one or more roads	4	4	87
101	shots of a basket being made - the basketball passes down through the hoop and net	2	4	104
102	shots from behind the pitcher in a baseball game as he throws a ball that the batter swings at		6	183
103	shots of Yasser Arafat	3		33
104	shots of an airplane taking off	1	2	44
105	shots of a helicopter in flight or on the ground	4	2	52
106	shots of the Tomb of the Unknown Soldier at Arlington National Cemetery	4		31
107	shots of a rocket or missile taking off Simulations are acceptable	4	4	62
108	shots of the Mercedes logo (star)	3		34
109	shots of one or more tanks	2	2	16
110	shots of a person diving into some water	3	1	13
111	shots with a locomotive (and attached railroad cars if any) approaching the viewer	3	4	13
112	shots showing flames	3	4	228
113	more shots with one or more snow-covered mountain peaks or ridges Some sky must be visible behind them	3	2	62
114	shots of Osama Bin Laden	3		26
115	shots of one or more roads with lots of vehicles	5	4	106
116	shots of the Sphinx	3		12
117	shots of one or more groups of people, a crowd, walking in an urban environment (for example with streets, traffic, and/or buildings)	4	4	665
118	shots of Congressman Mark Souder	2		6
119	shots of Morgan Freeman	3		18
120	shots of a graphic of Dow Jones Industrial Average showing a rise for one day The number of points risen that day must be visible		6	47
121	shots of a mug or cup of coffee	3	2	95
122	shots of one or more cats At least part of both ears, both eyes, and the mouth must be visible The body can be in any position	4	3	122
123	shots of Pope John Paul II	5	2	45
124	shots of the front of the White House in the daytime with the fountain running	2	3	10

Defined retrieval unit

A shot is defined in TRECVID2003 as a commonly agreed retrieval unit for system comparison Common shot boundary definitions were created and formatted in MPEG-7 by the CLIPS-IMAG group for distribution to all TRECVID2003 participants and this records the start time and shot duration

Indexing features available in TRECVID2003

Three types of indexing features are marked up m MPEG-7 format and shared among TRECVID2003 groups as listed below

- the output of an automatic speech recognition system transcripts donated by LIMSI [Gauvain et al, 2002] The ASR transcripts are presented as text annotations corresponding to each of the shots
- a closed-caption based transcript
- the output of 17 feature detectors donated by TRECVID2003 groups The feature detectors were defined by TRECVID2003 participants including outdoors, news subject face, people, building, road, vegetation, animal, female speech, car/truck/bus, aircraft, news subject monologue, non studio setting, sporting event, weather news, zoom-in, physical violence and person X (X specifically refers to Madeleine Albright) Their results show the presence and absence of the features within shots

Manual search submission requirements

Each manual system submission requires one baseline run to be submitted based only on the text from LIMSI ASR outputs For each topic, the system can return a ranked list of at most 1000 common reference shots from the search collection

7.2 A Review of Video Retrieval Techniques Used by TRECVID2003 Participants

TRECVID2003 Experiments continued on the study of combining different types of features in video retrieval, in particular the result fusion techniques and manual query formulation methods Table 7-3 below summarises the feature types used in

the manual search task by 6 TRECVID2003 participants CMU, Fudan, IBM, Lowlands, Imperial, Oulu/VTT

Table 7-3 Summary of features used in the manual search task by TRECVID2003 participants

Groups	Spoken document features	Concept-based features	Content-based features
Carnegie Mellon University	OKAPI formula, query expansion	All 17 features + their own (i e anchor and commercial)	Regional HSV colour, regional texture and regional edge direction histograms
Fudan University	Vector space model	All 17 features + their own	Colour histogram, edge direction histogram and face recognition
IBM Research	OKAPI formula	All 17 features plus their own (approx 70 concepts used)	Colour, texture, edges, shape, motion, etc
Lowlands Team	Language model	Anchor	A mixture of Gaussian models of DCT coefficients in the YCbCr colour space YcbCr color space Temporal information of shots was under consideration
Imperial College	Vector space model	None	Pre-processed all video images by removing the bottom 52-pixel lines where the news lines appear 8 different texture and colour features were extracted HSV colour histogram, colour structure and etc
University of Oulu/VTT	Vector space model	15 concepts were used as a query vector with binary values Concept "zoom-in" and "person X" were eliminated	Colour, Motion, and structure of edges

Result fusion techniques

A popular video retrieval system design among the TRECVID2003 participants can be summarised into the following steps (1) each shot is associated with a set of features such as ASR, content-based and concept-based features, (2) given different query types for a topic, a set of individual ranked lists from different feature search modules are returned, (3) finally these individual ranked lists are combined into one single ranked list based on a chosen fusion technique [Hauptmann et al, 2003] [Westerveld et al, 2003] [Heesch et al, 2003] [Rautianen et al, 2003]

The fusion technique is a linear weighing schema by setting weights based on query types For instance, Carnegie Mellon University grouped the 25 topics into two types topics on finding persons and topics on other non-persons [Hauptmann et al, 2003] For a topic on finding a person, the weights for the result fusion from four different search modules were set to (1) text = 2, (2) face =1, (3) colour = 1, (4) anchor =0 No optimal weights can be set during result combination and most TRECVID participants assigned weights subjectively [Westerveld et al, 2003] [Heesch et al, 2003] [Rautianen et al, 2003] Attempts were also made by CMU to learn the best weights via training labelled sets based on the TRECVID2003 training collection [Hauptmann et al, 2003] and their experiments showed that a run by manually setting weights was close to the one using learned weights

Manual query formulation methods

There were mainly three ways of formulating a query based on the query types used in TRECVID2003

- Manually selecting the keywords

  Most TRECVID2003 participants chose keywords that came from the text description for each topic A query expansion technique was applied to find more similar terms for each topic by CMU via learning similar terms based on the TRECVID2003 training collection [Hauptmann et al, 2003]

- Manually setting concept-based features

  Apart from the 17 concepts defined by TRECVID2003, participants could have used their own defined concepts to assist video retrieval CMU applied their anchor and commercial output detectors to exclude shots that contained anchor persons or commercials according to the requirement of specific topics [Hauptmann et al, 2003] Similarly, the Lowlands team implemented an anchor person detection algorithm to filter out anchor shots before further retrieval and their experiments have shown that a run by including anchor shots is very close to the one when excluding them [Westerveld et al, 2003]

- Automatically computing content-based features

  Various content-based features were applied in TRECVID2003 including colour, texture, edge direction histogram, etc Experiments were done to study the effect on performance by using multiple image examples in a query The Lowlands team and University of Oulu/VTT found the benefit of using carefully selected multiple image examples in a query [Westerveld et al, 2003] [Rautiamen et al, 2003]

  Further study of the temporal characteristics of video was conducted by the Lowlands team [Westerveld et al, 2003] They developed a dynamic retrieval model in which more than one key frame was sampled for each shot In comparison, a static retrieval model was defined as the one that only uses one key frame to represent each shot Their results have shown that the dynamic model slightly outperforms the static model

Table 7-4 lists the performance of the 34 submitted runs from 6 participants Each participant was required to provide a baseline run using only ASR transcripts for their own within-site comparisons as indicated in the table CMU experiments show that the inclusion of content-based and concept-based features results in higher Mean Average Precision (MAP) than a baseline run They considered the consistency of high MAP was caused by the small size of training data The Lowlands team found that if a given query works well in both text-only search module and content-based search module, a fusion of these two results can give better performance than either of the individual results

On the contrary, results by Imperial College indicate that the performance difference between a run that includes content-based features and one that excludes them was not significant enough to draw any conclusion Experiments from the University of Oulu/VTT also show that there was no improvement in performance over their baseline run when any combinations of content-based, concept-based and text features were used together

Table 7-4 Summary of performance of the TRECVID2003 manual search task

Group	Submitted Runs	Mean Average Precision	Total relevant shots returned	System Specification
CMU	CMU03	0 177	621	ASR (baseline)
	CMU04	0 207	846	ASR + concept-based +content-based
	CMU05	0 198	919	ASR + concept-based +content-based
	CMU06	0 218	862	ASR + concept-based +content-based
	CMU07	0 196	850	ASR + concept-based +content-based
	CMU08	0 198	805	ASR + concept-based +content-based
	CMU09	0 178	702	ASR + concept-based +content-based
Fudan	Fudan_1	0 035	180	Concept-based
	Fudan_2	0 055	92	ASR + concept-based
	Fudan_3	0 070	141	ASR + concept–based
	Fudan_4	0 062	228	ASR (baseline)
	Fudan_5	0 034	267	Concept -based + content-based (others)
	Fudan_6	0 018	322	Content-based (others)
	Fudan_7	0 043	330	Concept -based + content-based (colour only)
	Fudan_8	0 017	162	Content-based (colour only)
IBM	IBM-2	0 120	873	
	IBM-3	0 146	845	
	IBM-6	0 046	554	Concept-based or content-based
	IBM-7	0 043	386	Content-based
	IBM-8	0 085	626	
	IBM-9	0 123	873	
	IBM-10	0 090	539	
Imperial	ICL-1	0 074	219	ASR (baseline)
	ICL-2	0 076	430	ASR + content-based
Lowlands	LL11_ASR	0 130	719	ASR (baseline)
	LL11_dynamic	0 022	290	Content-based (temporal)
	LL11_dynabest	0 135	698	ASR + content-based (no anchor)
	LL11_Qmodel	0 005	140	Content-based (multiple images)
	LL11_staticML	0 022	276	Content-based (non-temporal)
Oulu	UOMT_M5	0 098	497	ASR (baseline)
	UOMT_M6	0 005	252	Content-based (One image)
	UOMT_M7	0 023	425	Content-based (Multiple image)
	UOMT_M8	0 024	477	Content-based (Multiple image) + ASR
	UOMT_M9	0 004	229	Content-based + concept-based (Multiple image)

For most of the topics, text-based queries give the best results and content-based queries might prove useful when there is no match between the query terms and index terms from the ASR However the usefulness of content-based features is limited due to the difficulty of searching a generic video collection using content-based features only A content-based search may produce good results where a topic is general such as a scene (i e baseball ground, basketball ground or Dow Jones graph) and bad results can be returned when the topic is too specific such as persons and things (i e tank and coffee cup)

Figure 7-1 shows the average precision by topic among the 34 runs. Topics 114 (Osama Bin Laden), 116 (Sphinx), 120 (Dow Jones graph) and 123 (Pope John Paul II) are among the top overall given the mean of top half of runs. For some topics such as 109 (tanks), 111 (locomotive approaching), 118 (Congressman Mark Souder) and 121 (a cup of coffee), it was found very difficult to find any relevant shots because the text-based matches were restricted and the searches mainly depend on content-based features.

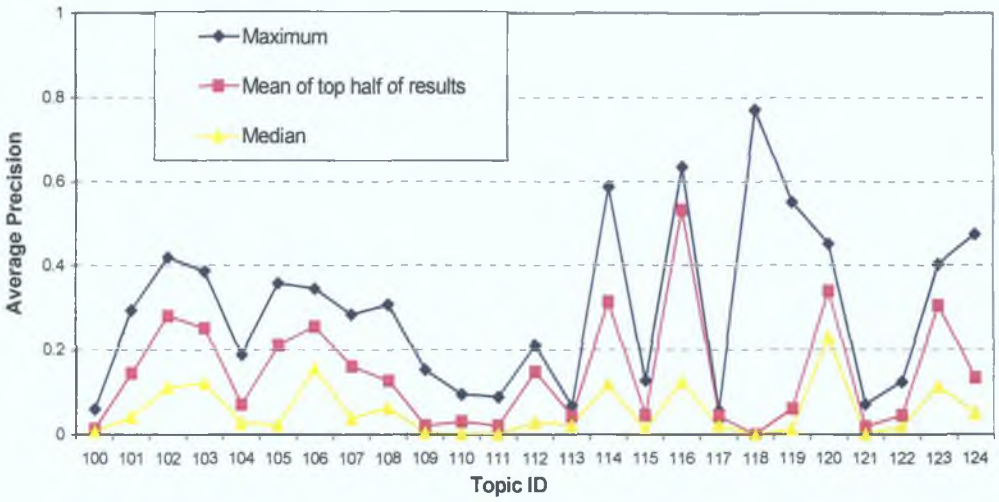


Figure 7-1: Manual search: average precision per topic

7.3 Experimental Settings

Spoken-text, concept-based and content-based features are major components in video retrieval. The results obtained in the TRECVID2003 experiments again indicated the important role of semantically rich text-based search in video retrieval and with selective inclusion of content/concept-based features it is possible to give better performance. We carried out our experiments on the TRECVID2003 collection to study the same four issues as listed in section 6.3.1.

The experimental settings are described in this section which details the construction of a baseline model and various modified systems for comparison Our experimental results will be given in section 7 4

Our baseline systems

A conventional video retrieval model using the LIMSI ASR text outputs only was implemented as a baseline system The indexing model used was a Vector Space model adopting the TF*IDF term weighting algorithm The ASR transcript that belongs to each of the shots was solely contributing to generating a term weight vector The steps for constructing the indexing and retrieving relevant shots were given in section 6 3 1

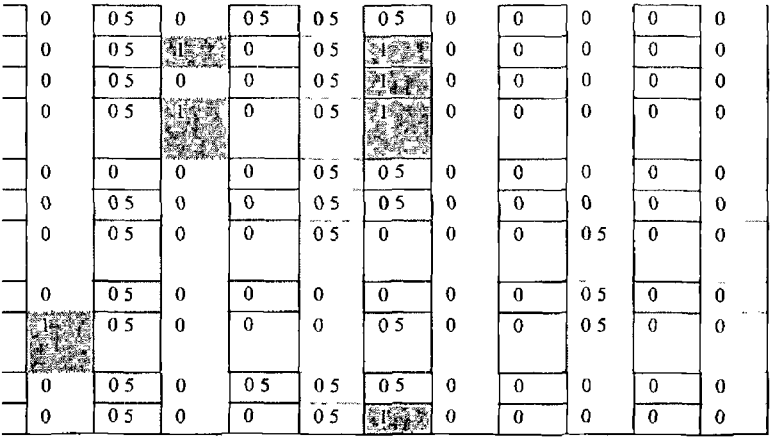
Query preparation

We prepared different types of queries for TRECVID2003 in a similar way to that for TRECVID2002 Query terms for each topic were taken strictly from the topic description (see Table 7-5 below), to form a text query We assigned a value of 0, 0 5 or 1 to each concept defined by NIST to construct a concept-based query and the query had a length of 17 (see Table 7-5) Finally a content-based query was obtained by automatically computing three features for each image example (1) four region * 16 bin scalable colour (i e colour histogram), (2) nine region * dominant colour using the RGB colour space (the number of dominant colour for each RGB colour component is 2), (3) 80 bin edge component histogram

Table 7-5 Selected text query and manually created concept-based query vector by topic in TRECVID2003

Topic ID	Text Query	Concept-based query																
		Out-doors	News-Subject Face	People	Building	Road	Vegetation	Animal	Female Speech	Car Truck Bus	Aircraft	News Subject Monologue	Non-Studio Setting	Sporting Event	Weather News	Zoom- in	Physical Violence	Person X
100	Aerial views containing buildings roads		0	0			0	0	0 5	0	0 5	0 5		0	0	0	0	0
101	Basket made basketball passes down through hoop net	0	0		0	0	0	0	0 5	0	0	0 5			0	0	0	0
102	Behind pitcher baseball game throws ball batter swings		0		0	0	0	0	0 5	0	0	0 5			0	0	0	0
103	Yasser arafat	0 5		0	0	0	0	0	0 5	0	0	0 5	0 5	0	0	0 5	0	0
104	Airplane taking off		0	0	0	0	0	0	0 5	0		0 5		0	0	0 5	0	0
105	Helicopter flight ground		0	0	0 5	0 5	0	0	0 5	0		0 5		0	0	0	0	0
106	Tomb unknown soldier Arlington national cemetery		0	0	0	0	0	0	0 5	0	0	0 5		0	0	0 5	0	0
107	Rocket missile taking off simulations		0	0	0	0	0	0	0 5	0	0	0 5		0	0	0	0	0
108	Mercedes logo	0	0	0	0	0	0	0	0 5	0	0	0 5	0	0	0	0 5	0	0
109	Tanks	0 5	0	0	0 5	0 5	0 5	0	0 5	0	0	0 5		0	0	0	0	0
110	Person diving water	0 5	0	0 5	0	0	0	0	0 5	0	0	0 5			0	0	0	0
111	Locomotive railroad cars approaching viewer		0	0	0	0 5	0	0	0 5	0	0	0 5		0	0	0	0	0
112	Flames	0 5	0	0	0 5	0	0	0	0 5	0 5	0 5	0 5		0	0	0	0	0
113	Snow covered mountain peaks ridges sky visible		0	0 5	0	0	0 5	0 5	0 5	0	0	0 5		0	0	0	0	0

114	Osama Bin Laden	0 5	1	0	0	0	0
115	Roads with lots of vehicles	1	0	0	0 5	1	0
116	Sphinx	1	0	0	0	0	0
117	People crowd walking urban environment streets traffic buildings	1	0	1	1	1	0
118	Congressman Mark Souder	0 5	1	0	0	0	0
119	Morgan Freeman	0 5	1	0	0	0	0
120	Dow Jones Industrial Average showing rise number points risen day visible	0	0	0	0	0	0
121	Mug or cup of coffee	0	0	0	0	0	0
122	Cats at least part ears eyes mouth visible body position	0 5	0	0	0	0	0
123	Pope John Paul	0 5	1	0	0	0	0
124	Front White House daytime fountain running	1	0 5	0	0 5	0	0



Our submission

For each topic, each our system returned a ranked list of 100 common reference shots from the search collection

Evaluation Systems

We designed seven systems for evaluation in the same way to that of our TRECVID2002 systems (for details see section 6 3 1) Table 7-6 below lists the evaluation systems based on the selection of index types and query types "Sys1" is the baseline system which implements the conventional TF*IDF model The rest of the systems incorporate content-based/concept-based features in both the indexing and retrieval process accordingly

Table 7-6 Evaluation system design for TRECVID2003

		Methods of formulating a query for a given topic				
		QW_j	$QW_j^{content}$	$QW_j^{concept}$	$QW_j + QW_j^{content}$	$QW + QW_j^{concept}$
Methods of index aggregation	SW_{ij}	Sys1				
	SW_{ij} + $DerivedSW_{ij}^{content}$	Sys2	Sys3		Sys4	
	SW_{ij} + $DerivedW_{ij}^{concept}$	Sys5		Sys6		Sys7

Variable Settings

From our previous experiments in TRECVID2002, we have learned that there is no performance difference in introducing two variables during clustering the window size *WIN* and the threshold *xT* For our experiments on the TRECVID2003 collection we now limit *WIN* to 0 and 1 and the *xT* setting remains as 0 1, 0 15 and 0 2 but only applies to *WIN*=1 Also it has been shown in Chapter 6 that varying variable *topK* (the number of most similar clusters chosen to map a non-text query onto a query term vector) did not given any performance benefit The higher the *topK* value is the worse performance a system can have We therefore set *topK* to 1, 3 and 5

We introduce a new variable noted as *N* to replace variable *K* *K* is originally the number of clusters that are required in the k-means clustering module for each news programme Variable *N* accounts for the number of shots within a cluster and can be used to calculate the number of clusters *K* Since the total number of shots of each

news programme varies, instead of setting different K values in each run, we assigned the same value to N (the number of shots within a cluster) for all programmes N is chosen to be 5, 7, 9 and 11

7.4 Experiments

Section 7 4 1 will examine four systems and discuss their relative performances when using a combined index built from concept-based and content-based features in conjunction with spoken-text features Whether a combined query inferred from a non-text query along with the original text query will give reasonable performance will be illustrated for each of the 3 systems in section 7 4 2 Also discussed is a comparison of the performance difference between using one best image example and using all image examples in a combined query The comparison will take Sys4 as an example to consider content-based features only in section 7 4 3 Finally, back to the analysis of non spoken-text features in video retrieval, systems Sys3 and Sys6 will look into the performance of using queries which are content-based and concept-based alone through a combined index, respectively

7.4.1 Retrieval Performance When Using an Aggregated Index

The analysis of results illustrated by Sys1, Sys2 and Sys5 can help us understand the effects of including content/concept-based features in a traditional TF*IDF index for the TRECVID2003 search collection All systems used text only queries as given in Table 7-5

- Sys1 our baseline system using a traditional TF*IDF index,
- Sys2 an aggregated index by incorporating content-based features into the traditional TF*IDF index,
- Sys5 an aggregated index by incorporating concept-based features into the traditional TF*IDF index

Tables 7-7 and 7-8 below summarise the mean average precision by variable *N*, *WIN* and *xT* for Sys2 and Sys5, respectively No major performance difference was found by varying the parameters in both systems

Table 7-7 Mean average precision by variable N, WIN and xT for Sys2

N	WIN=0	WIN = 1		
		xT=0 1	xT=0 15	xT=0 2
5	0 0529	0 0574	0 0548	0 0454
7	0 0586	0 0457	***0 0599***	0 0465
9	0 0542	0 0445	0 0562	0 0477
11	0 0574	0 0572	0 0448	0 0588

Table 7-8 Mean average precision by variable N, WIN and xT for Sys5

N	WIN=0	WIN = 1		
		xT=0 1	xT=0 15	xT=0 2
5	0 0562	**0 0634**	0 0616	0 0515
7	0 0548	0 0581	0 0606	0 0529
9	0 0545	0 0542	0 0537	0 0533
11	0 0578	0 0554	0 0594	0 0533

We took the best results produced by Sys2 and Sys5 in order to compare the performance of our baseline system Sys1 MAP has a peak for Sys2 when *N*=7, *WIN*=1 and *xT*=0 15 The best MAP for Sys5 was given when *N*=5, *WIN*=1, *xT*=0 1 Table 7-9 illustrates a comparison of the overall performance among the three systems

Table 7-9 Summary of performance of Sys1, Sys2 and Sys5

System	MAP	Total relevant shots retrieved
Sys1	0 0594	163
Sys2	0 0599	169
Sys5	0 0634	164

Figure 7-2 below shows the precision at 11-recall levels and Figure7-3 gives the precision at 6-document cut-off levels for Sys1, Sys2 and Sys5 All three systems are very comparable to each other It shows that an aggregated index that includes content/concept-based features into a traditional TF*IDF index provides no overall

improvement in performance over the baseline system in the TRECVID2003 search collection.

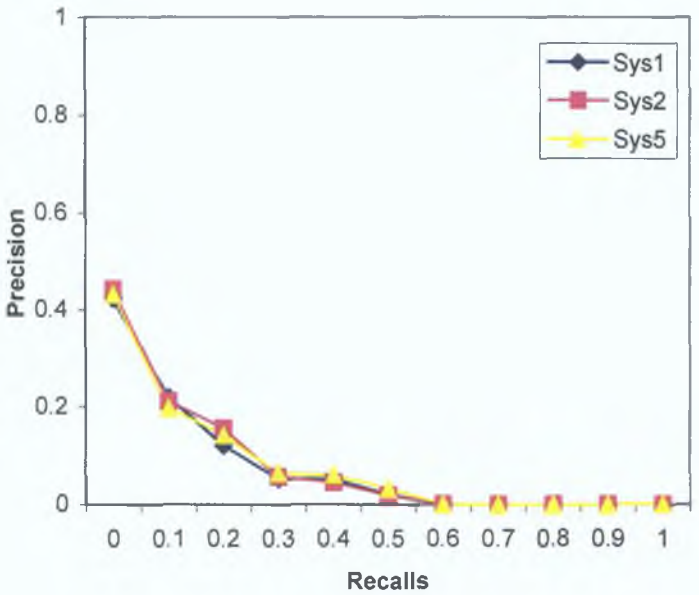


Figure 7-2: Precision at recalls for Sys1, Sys2 and Sys5

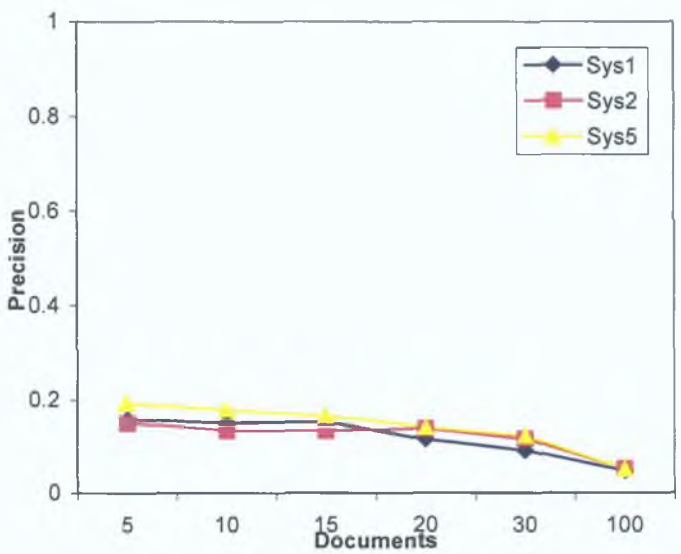


Figure 7-3: Precision at document cut-offs for Sys1, Sys2 and Sys5

Figure 7-4 illustrates the average precision over the 25 topics for Sys1, Sys2 and Sys5 and the median average precision (noted as TREC_Median) by TRECVID2003 participants. Except for topics 114 (Osama Bin Laden) and 116 (Sphinx), all systems

worked very comparably. Sys2 provides slightly higher average precision for topic 114 while Sys5 gives better performance for topic 116. Performance improvement was gained by using an aggregated index over certain topics to a small degree.

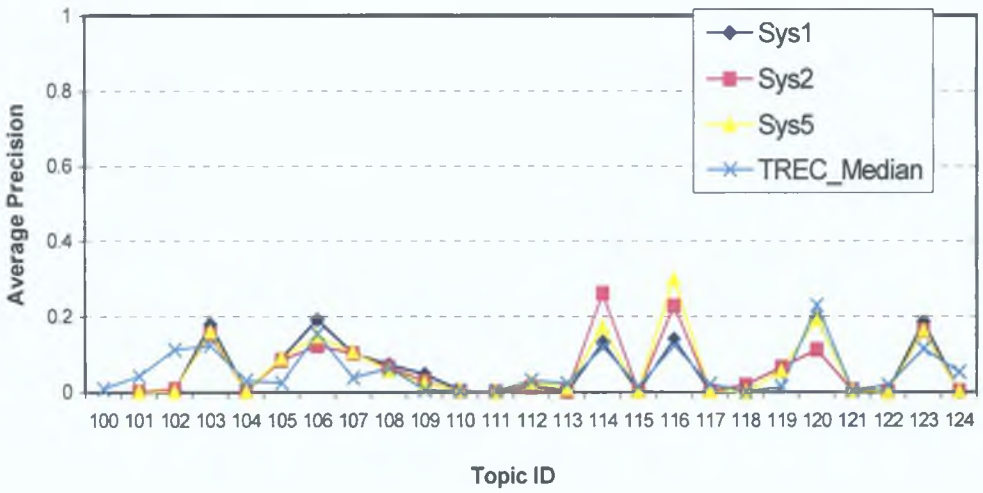


Figure 7-4: Average precision by topic for Sys1, Sys2, Sys5 and TREC_Median

7.4.2 Retrieval Performance When Using an Aggregated Query

This section shows the results of using an aggregated query which includes non-text queries into a text query in video retrieval. We will first study Sys2 and Sys4 to see the influence of inclusion of content-based queries. Sys5 and Sys7 will then be analysed for inclusion of concept-based queries.

Sys2 vs. Sys4

Sys2 is the baseline system for Sys4. Both systems use the same aggregated index which is derived from content-based features and ASR output texts. The only difference between them is seen in the query formulation process: Sys2 accepts text-only queries while Sys4 uses aggregated queries including content-based queries. Only one image example was manually selected to generate a content-based query for each topic.

Previous study of variable *N*, *WIN* and *xT* in section 7 4 1 shows that there is no major difference among the settings of variables We chose the aggregated index of *N*=7 and *WIN*=0 from Sys2 (see Table 7-7 where MAP has a value of 0 0586) as the index for Sys4 Two variables are under consideration in Sys4 (1) the number of most similar clusters chosen to map a content-based query onto a query term vector – *topK*, (2) the weight *PF* given to an original text query during the creation of an aggregated query Variable *topK* is set to 1,3 and 5 Weight *PF* is set to 20, 25 and 30

Table 7-10 below gives the mean average precision by variable *topK* and *PF* for Sys4 The performance drops slightly as variable topK increases As more similar clusters are found, the number of cluster terms unrelated to a topic that are used as query terms increases

A slight increase was found in MAP as parameter *PF* increases MAP has the same value as that of Sys2 when *PF*=25 and *topK*=1 and gives its highest value to 0 0594 Variable *PF* has shown its ability to balance the influences from the importance of original query terms and from the inclusion of extra query terms obtained from content-based queries

Table 7-10 Mean average precision by variable topK and PF for Sys4 using one best image example in a query

TopK	Weight PF		
	20	25	30
1	0 0584	0 0586	**0 0594**
3	0 0553	0 0556	0 0561
5	0 0540	0 0542	0 0547

Sys5 vs Sys7

Sys5 is the baseline system for Sys7 Both systems use the same aggregated index which is derived from concept-based features and ASR output texts The only distinction between them is that Sys5 handles text-only queries while Sys7 deals with aggregated queries which includes concept-based queries

Previous study of variable *N*, *WIN* and *xT* in section 7 4 1 shows that there is no major difference among the settings of variables We chose the aggregated index of *N*=11 and *WIN*=1 and *xT*=0 15 from Sys5 (see Table 7-8 where MAP has a value of 0 0594) as the index for Sys7

Table 7-11 Mean average precision by variable topK and PF for Sys7

TopK	Weight PF		
	20	25	30
1	0 0524	0 0567	**0 0580**
3	0 0436	0 0455	0 0474
5	0 0398	0 0417	0 0437

Table 7-11 above gives the mean average precision by variable *topK* and *PF* for Sys7 The performance drops slightly as variable topK increases A slight increase was found in MAP as parameter *PF* increases MAP has its highest value at 0 0580

Our previous study on the TRECVID2002 collection illustrated that concept-based queries were too general to find any similar cluster for query derivation The result shows that using an aggregated query that come from a text-only and concept-based query for video retrieval can maintain the same performance level as using a text-only query by assigning high values to variable *PF*

Sys2, Sys4, Sys5 vs Sys7

We took the best results of Sys4 and Sys7 for comparison with their corresponding baseline systems Sys2 and Sys5 Table 7-12 lists the performance of each system It is seen that all four systems perform very similarly There is no significant improvement over the usage of an aggregated query for video retrieval in the TRECVID2003 search collection It is due to the difficulty of finding the correct clusters to a non-text query for mapping onto a text description

Table 7-12 Summary of the performance of Sys2, Sys4, Sys5 and Sys7

	Content-based features		Concept-based features	
	Sys2	Sys4	Sys5	Sys7
MAP	0 0586	0 0594	0 0594	0 0580
Total relevant shots retrieved	155	159	166	163

Figures 7-5 and 7-6 below plot the precision at 11-recall levels and at 6-document cut-off levels for Sys2, Sys4, Sys5 and Sys7, respectively. Sys7 produced lowest precision at recall level of 0 and at document level of 5.

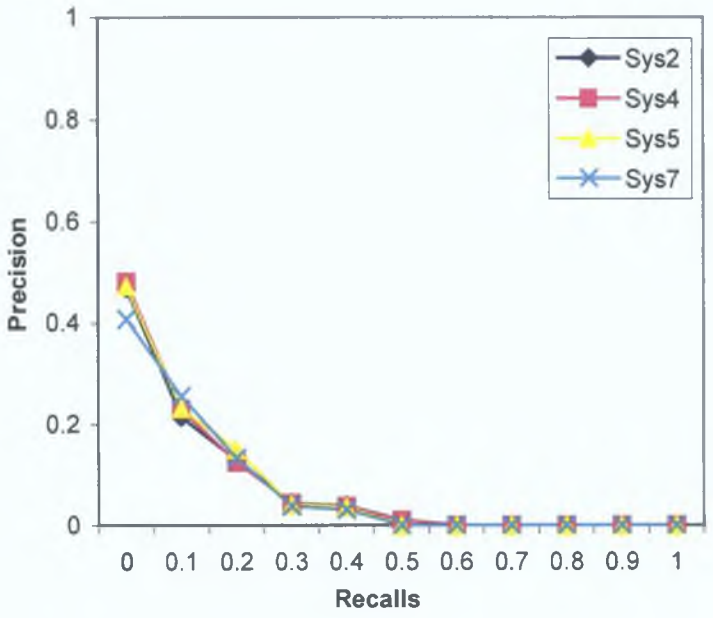


Figure 7-5: Precision at recalls for Sys2, Sys4, Sys5 and Sys7

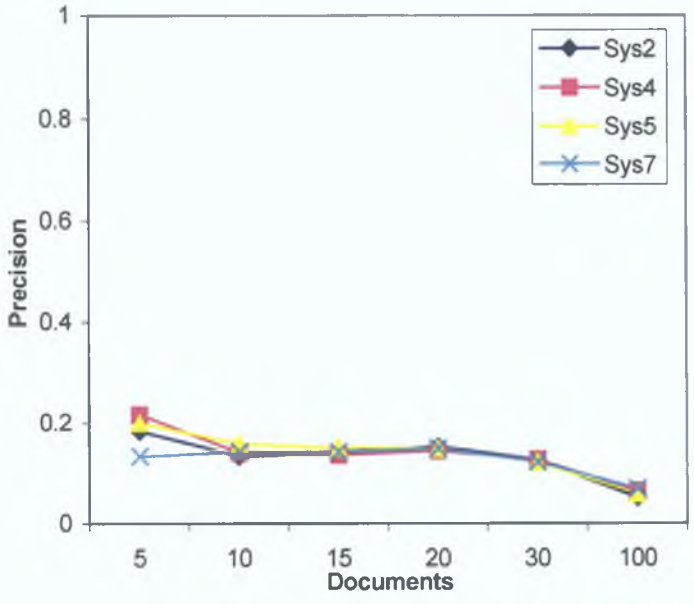


Figure 7-6: Precision at document cut-offs for Sys2, Sys4, Sys5 and Sys7

Figure 7-7 gives the average precision by topic over the four systems. Interestingly, Sys7 did well in topic 118 (Congressman Mark Souder), in which a concept-based query was used to map onto a text description.

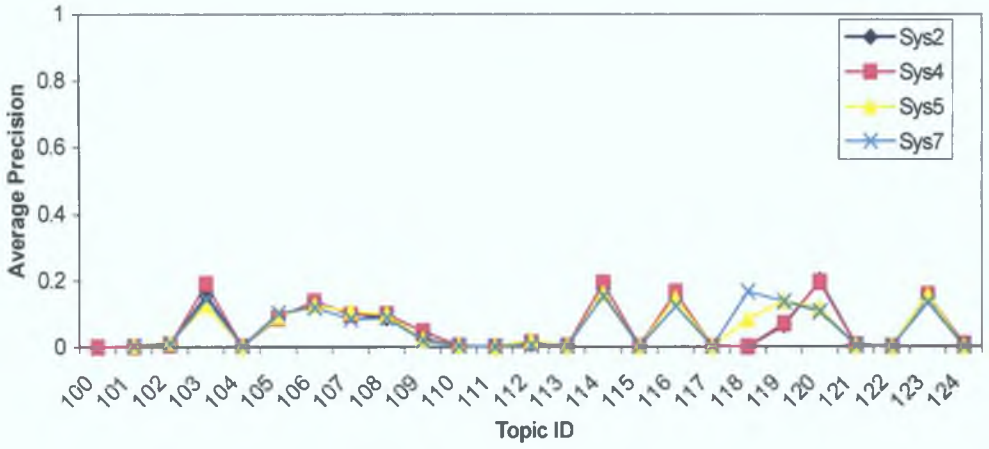


Figure 7-7: Average precision by topic for Sys2, Sys4, Sys5 and Sys7

7.4.3 Retrieval Performance When Using Multiple Image Examples in a Query

Similarly to section 6.4.3, we are interested in studying the performance difference between using one best image example and using all image examples in a query for the TRECVID2003 search collection. Only content-based queries are under consideration. Sys4 then has two variations. The first variation noted as Sys4_1 is to use only one best image example in a query. The chosen image example should be very similar to the search collection regarding the picture quality. The results from Sys4_1 are listed in Table 7-10 in section 7.4.2.

The second variation is noted as Sys4_2 and is to use all image examples in a query for each topic. Table 7-13 below shows the mean average precision by variable *TopK* and *PF* for Sys4_2. In comparing Table 7-10 and Table 7-13, it can be seen that Sys4_1 performs better than Sys4_2 over all variable settings. This poorer result for adding more than one image in a query example is due to the increased noise when using many image examples in a query. There are twice more image examples per topic provided in TRECVID2003 than those in TRECVID2002.

The more image examples used, the more incorrect clusters that are found for a content-based query mapping to a text description

Table 7-13 Mean average precision by variable TopK and PF for Sys4_2

TopK	Weight PF		
	20	25	30
1	0 0492	0 0509	**0 0514**
3	0 0381	0 0435	0 0465
5	0 0289	0 0345	0 0368

Table 7-14 below summarises the best performance of Sys4_1 and Sys4_2 Figure 7-8 and Figure 7-9 below plot the precision at 11-recall levels and at 6-document cut-off levels Figure 7-10 graphs the average precision by topic for Sys4_1 and Sys4_2 It can be seen that Sys4_1 outperforms Sys4_2 suggesting that not all image examples provided should be used in a query Careful selection of image examples is required for video retrieval in the TRECVID2003 search collection in order to get better retrieval performance

Table 7-14 Summary of performance of Sys4_1 and Sys4_2

	Sys4_1	Sys4_2
MAP	0 0594	0 0514
Total relevant shots retrieved	159	150

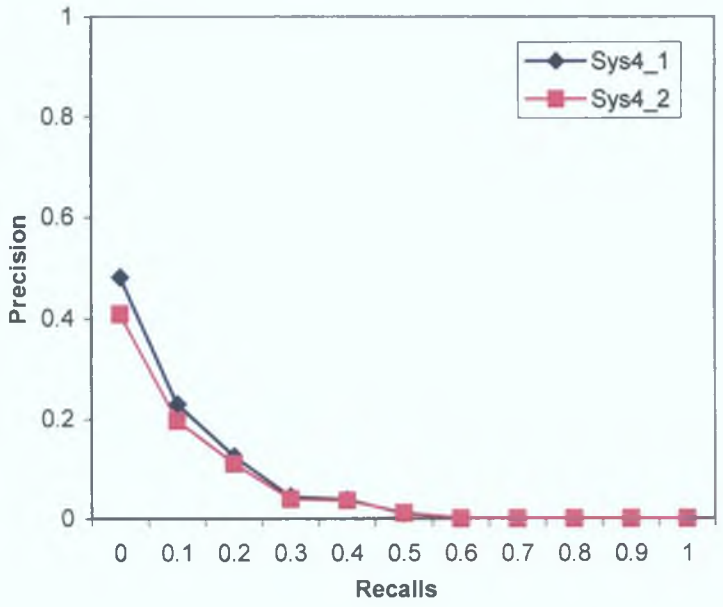


Figure 7-8: Precision at recalls for Sys4_1 and Sys4_2

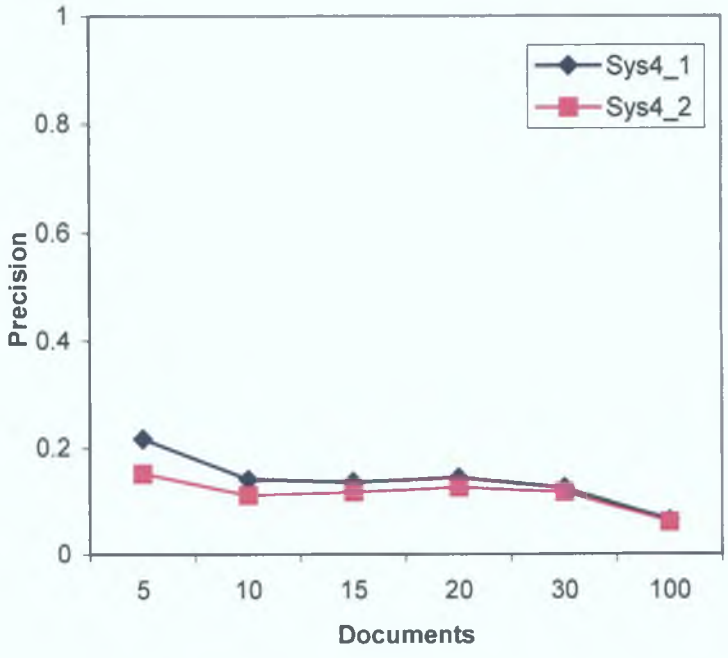


Figure 7-9: Precision at document cut-offs for Sys4_1 and Sys4_2

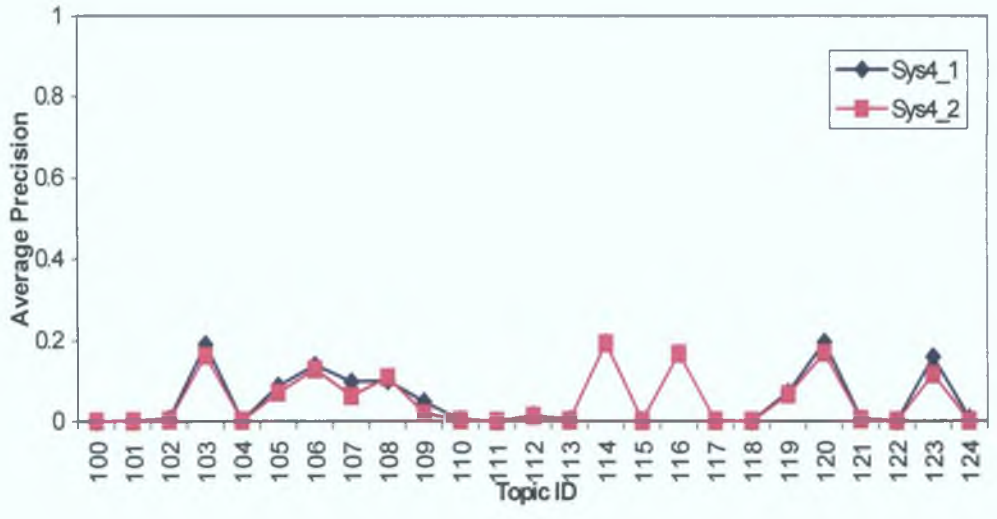


Figure 7-10: Average precision by topic for Sys4_1 and Sys4_2

7.4.4 Retrieval Performance When Using non-Spoken Text Features

Finally, additional experiments were carried out to study the performance difference when using a content/concept-based query alone for video retrieval in the TRECVID2003 collection. Sys3 and Sys6 are of interest here:

- *Sys3*: an aggregated index is applied that includes content-based features in a conventional TF*IDF index. An aggregated query is used that is derived from a content-based query. We only consider one best image example in a query in the query formulation process.
- *Sys6*: an aggregated index is applied that comprises concept-based features as part of the conventional TF*IDF index. An aggregated query is employed that is obtained from a concept-based query.

Table 7-15 and Table 7-16 below give the mean average precision by variables *topK* and *N* for Sys3 and Sys6, respectively. Sys3 performs better than Sys6 over most variable settings.

Table 7-15 Mean average precision by variable topK and N for Sys3

TopK	N			
	5	7	9	11
1	**0 0180**	0 0096	0 0170	0 0150
3	0 0051	**0 0180**	0 0156	0 0103
5	0 0042	0 0151	0 0147	0 0038

Table 7-16 Mean average precision by variable topK and N for Sys6

TopK	N			
	5	7	9	11
1	0 0036	0 0034	0 0025	0 0056
3	0 0044	0 0029	0 0028	**0 0066**
5	0 0045	0 0033	0 0039	0 0050

We choose the best performance for Sys3 and Sys6 and summarise these in Table 7-17 MAP for Sys6 is very low at 0 0066 and there are only 7 relevant shots retrieved over all topics Figure 7-11 and Figure7-12 graph the precision at 11-recall levels and at 6-document cut-off levels for both systems, respectively It is seen that the mapping of a non-text query to a text description is challenging, in particular for a concept-based query

Table 7-17 Summary of performance of Sys3 and Sys6

	Sys3	Sys6
MAP	0 0180	0 0066
Total relevant shots retrieved	68	7

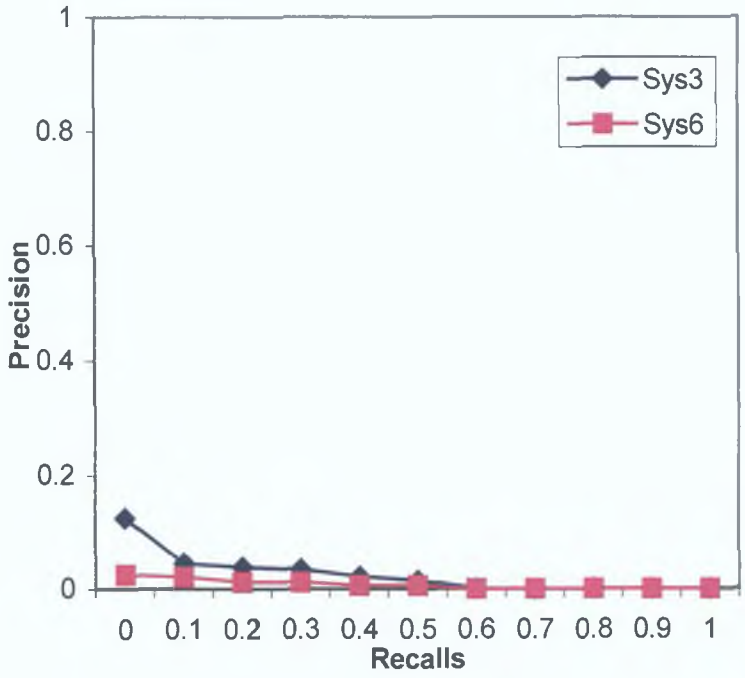


Figure 7-11: Precision at recalls for Sys3 and Sys6

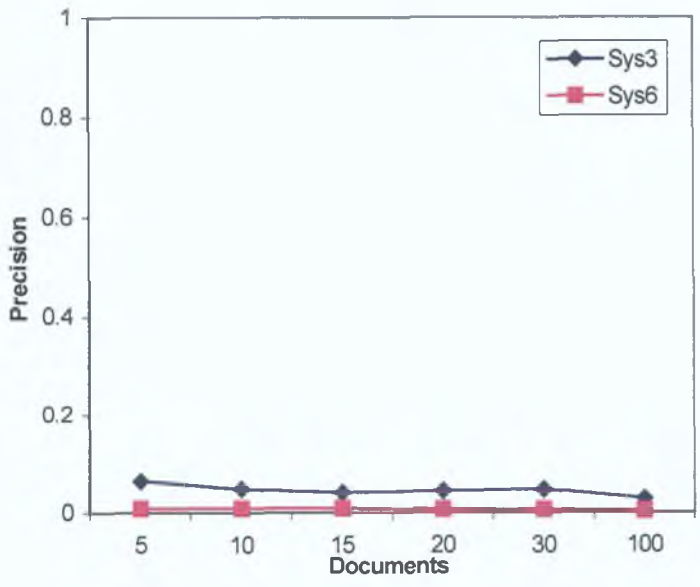


Figure 7-12: Precision at document cut-offs for Sys3 and Sys6

Figure 7-13 plots the average precision over the 25 topics for Sys3 and Sys6. Both systems retrieved a small amount of relevant shots for a limited number of topics. Sys3 did well in topic 118 (congressman Mark Souder) and topic 120 (Dow Jones Industrial graph).

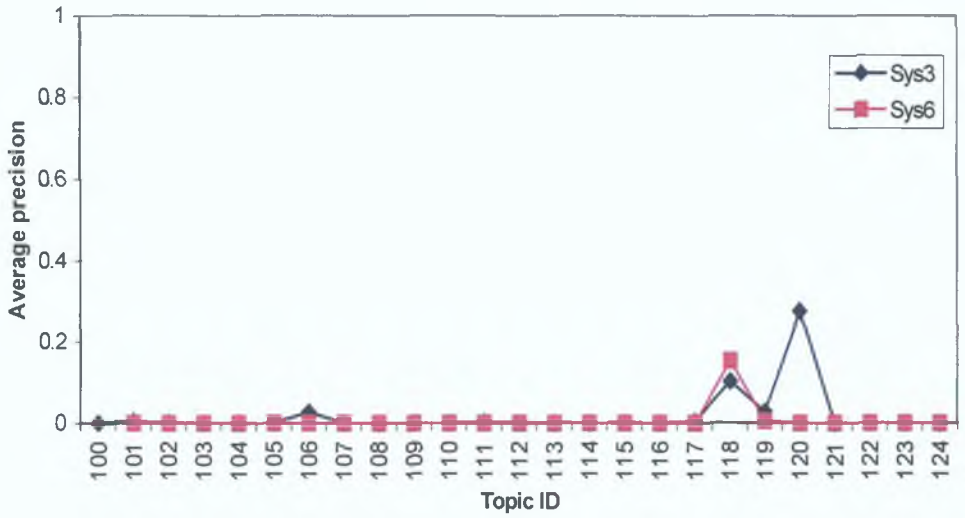


Figure 7-13: Average precision by topic for Sys3 and Sys6

7.5 Discussion of our TRECVID2003 Experiments

Our TRECVID2003 experiments have shown that there is no improvement in performance over our baseline run when any combination of content-based, concept-based and text features are used together.

The TRECVID2003 collection contains primarily TV news programmes. The typical structure of a TV news programme consists of news headlines and stories. News headlines are presented in the beginning of the programme to introduce the major stories to come and are essentially an abstract of the content. The most important news is delivered first, and less important news follows. Each story is self-contained and has three basic parts: the opening, in which an anchor person tells the story quickly; the narrative, a chronological retelling of events by showing some related shots; and the commentary, a conclusion/justification of the story by a corresponding news reporter.

There are few shots in the narrative part that are similar visually and semantically for each single news programme Clustering shots within a single news programme is thus ineffective compared to clustering within a programme in the TRECVID2002 collection In fact, adding semantic meaning to a shot from its neighbour shots in a video may seem more effective than adding meaning from clusters

For the important news background information to the story might also be placed as part of the narrative in such a way that shots in previous news programmes might be re-used In this case, clustering shots for a number of news programmes in succession may find more shots that are similar visually and semantically for clusters We redid our experiments by comparing a system using an aggregated index with our baseline system Sys1 The system for comparison noted as SysChron is constructed as follows

We first separate the TRECVID2003 videos into 3 types CNN, ABC and C-SPAN Since C-SPAN videos are about government committee meetings we handle them as before in section 7 4 1 and cluster shots for one C-SPAN video at the video level CNN and ABC news programmes are processed in five steps

1 Sort CNN and ABC videos in chronological order, respectively

2 For shot clustering we applied a sliding window approach on X CNN (/ABC) videos at the one time, where X is the number of videos after the current video for clustering (including the current one) Figure 7-14 illustrates a 3-video window for clustering shots from video 3 and its successive videos 4 and 5 Shots are clustered based on content-based features

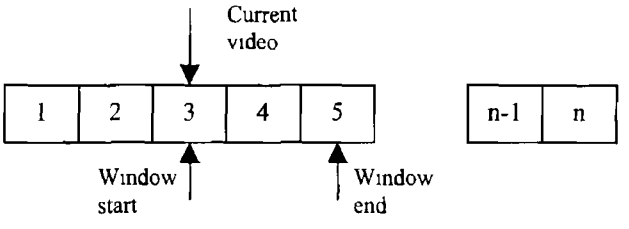


Figure 7-14 A 3-video sliding window

3 Assign meanings to each cluster using a modified TF*IDF algorithm as shown in Formula 6-3

4 Infer the meanings of shot based on cluster meanings A shot could fall into more than one cluster because each video is used in clustering for at most X times The weight of token W_j in a given shot S_i derived from the associated cluster C_k, $DerivedSW_{ij}^{k}$, is shown in Formula 6-4 We then aggregate the term weight $DerivedSW_{ij}^{k}$ from the X clusters that are related to shot S_i, as given in Formula 7-1

$$InferredSW_{ij} = \underset{k=1}{\overset{k\le X}{Agg}}(InferredSW_{ij}^{k}) \qquad (7\text{-}1)$$

5 Aggregate the inferred term weights with the traditional TF*IDF term weights to produce a single final index for video retrieval

Table 7-18 gives the mean average precision by variable X for Sys1 and SysChron All systems dealt with only text queries When we look at the four SysChron systems the MAPs remain similar at around 0 060 when X used =3, 5 and 7 videos The MAP drops to 0 052 when X increases to 9 The more videos involved in clustering, the more difficult it is to group shots that are similar semantically This is because each time a video added into the clustering process, a very small portion of shots that are semantically similar is introduced, and a very large portion of dissimilar shots also included Thus the larger the number of videos included in the clustering process the more distorted the cluster meanings

We compare the best performance of SysChron (X=3) with our baseline system Sys1. Both MAPs are very similar but SysChron retrieves less relevant shots over all 25 topics. There is no performance benefit over the baseline system when using more than one successive video for clustering.

Table 7-18: Mean average precision by variable X for Sys1 and SysChron

System	Sys1	SysChron (when X =)			
		3	5	7	9
MAP	0.0594	0.0624	0.0596	0.0613	0.0524
Total relevant shots returned	163	147	161	158	138

Figures 7-15 and 7-16 plot the precision at 11 recall levels and at 6 document cut-off levels for Sys1 and SysChron, respectively. Both graphs show that these two systems perform very similarly. Figure 7-17 gives the average precision by topic for both systems. In comparison with the baseline system Sys1, system SysChron did well in topics 118 ("Mark Soulder") and 116 ("Sphinx"), but performed poorly in topic 106 ("tomb of the unknown soldier") and 120 ("Dow Jones Industrial Graph").

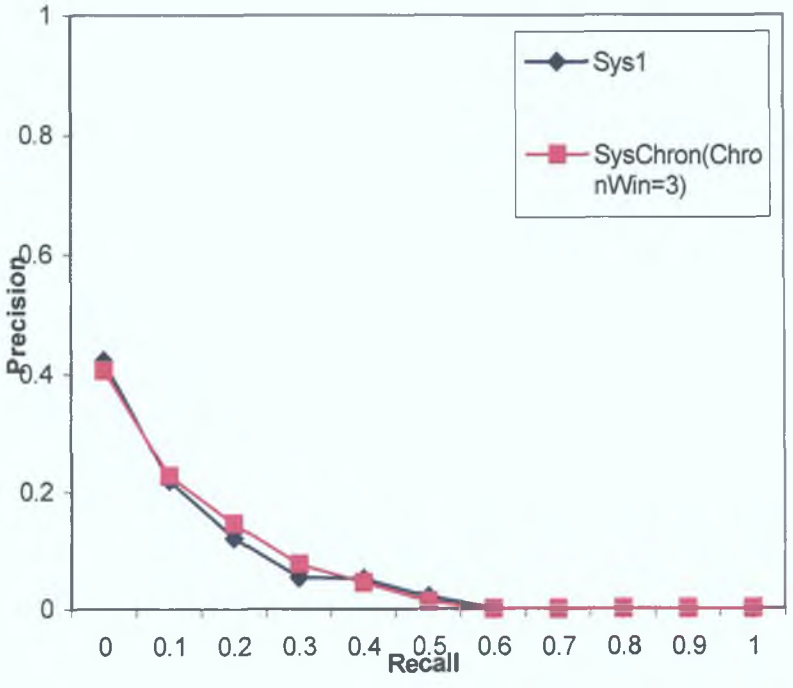


Figure 7-15: Precision at recalls for Sys1 and SysChron

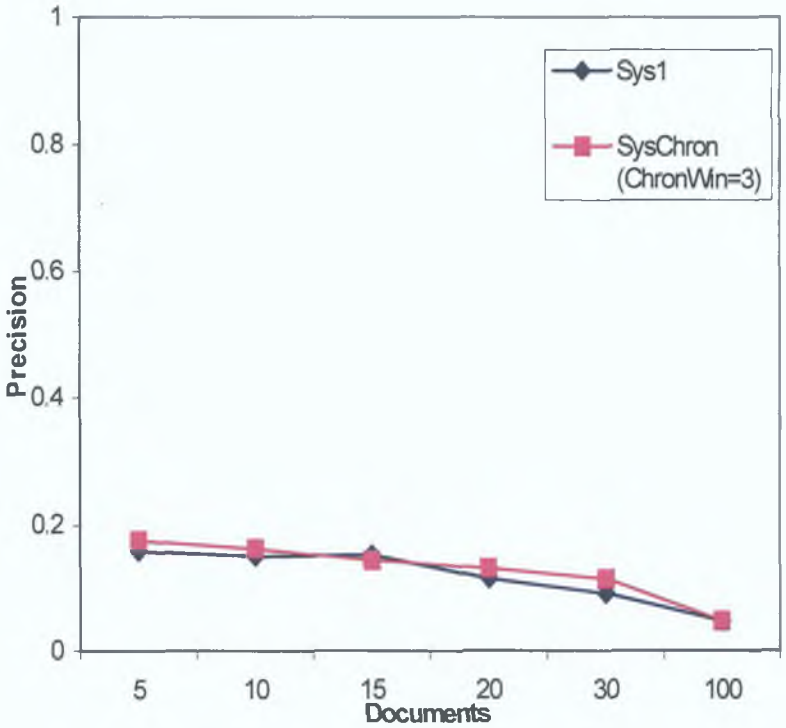


Figure 7-16: Precision at document cut-offs for Sys1 and SysChron

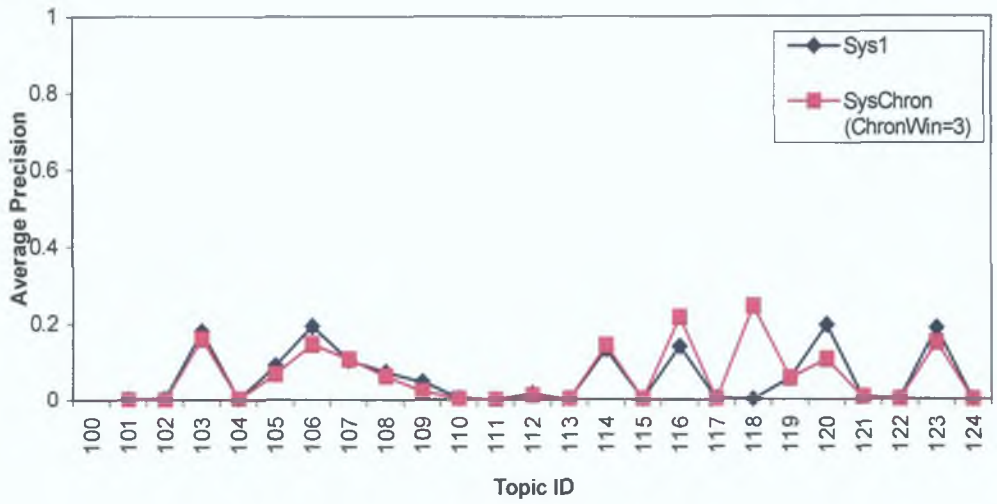


Figure 7-17: Average precision per topic for Sys1 and SysChron

7.6 Conclusions

Our TRECVID2003 experiments have shown that there is no significant improvement on employing an aggregated index when taking in content/concept-based features for the index creation These outcomes give the opposite position to the TRECVID2002 experiments in which performance improvement was found to some extent The main reason for the discrepancy might be different genres of the two search collections being studied

The TRECVID2002 collection has videos of documentaries from the 1940s to 1960s without commercials The characteristic of the old type of documentaries is that there is a good amount of shots within a video which are unnecessarily the same but similar visually and semantically In other words, the content is usually semantically similar for each video Spoken audio is clear with good mapping to what are shown on screen generally speaking, unlike today's video content and so it turns out to be an advantage for our assumption that shots in the same cluster are not only similar visually but also semantically to certain extent

The TRECVID2003 collection contains mainly videos of CNN and ABC broadcast TV news programmes The news programmes were blended with a number of different news stories, commercials, sports and weather Except for anchor person, commercial and sport shots, there are few shots in the remainder that are similar visually and semantically for each single news programme, but it is possible to encounter a good number of them across all broadcast news programmes Clustering shots within a single news programme may thus seem ineffective compared to clustering within a programme in the TRECVID2002 collection

Experiments also have been carried out one to clustering shots for more than one successive news programme to create the final index and no performance benefit was found over our baseline system There are shots that are similar visually and semantically across a few news programme unfortunately the total number of them is very small in comparison to the number of dissimilar shots There is chance of putting dissimilar shots into the correct cluster than similar ones The meaning of the cluster would be altered as more dissimilar shots are added to it

In the query formulation process, mapping a content-based query onto a text description and combining it with the original text query for video retrieval is topic-specific The challenge of finding the correct clusters for the mapping process remains The mapping of a concept-based query is less possible due to the generality of the defined concepts and duplicate formulation of different queries

Our TRECVID2003 experiments have also shown that using multiple image examples in a query is not as effective as using one single best image example The more dissimilar examples that are fed into the query formulation module at the one time, the more query terms are introduced that are unrelated to a given topic As a result of this, the unrelated terms can degrade the system performance

Chapter Eight

Conclusions

Digital video retrieval has developed as a branch of information retrieval – the science of designing systems to store items of information that need to be indexed, searched and shared to various user populations Considering the heterogeneous nature of video data, developing methods of digital video retrieval is even more of a challenge than the same process for other media types Text and content-based image retrieval methods can be extended to video retrieval, but such extensions are not straightforward Video is a structured document in which objects, actions and events are assembled to tell stories or convey particular visual information

Considering a video as a single document, indexing text features alone will give relatively high probability of a video being discriminated from the others, but will not be able to locate stories in a long video sequence Considering a video as a sequence of frames, indexing each of them as a still image will help find the visual information of interest, but lead to high redundancy To improve digital video retrieval, therefore, we need to identify structures of video by decomposing video into basic components and then create indices based on the structural information as well as the information from the representative frames

This thesis begins with an introduction and a description of the principle components of a video retrieval system, including the representations of video content and structure, indexing methods, video similarity and query formulation and visualisation To obtain these components, three major processes are normally involved video parsing, feature extraction, and abstraction Parsing is the detection of shot boundaries and identification of stories or scenes of videos Feature

extraction is similar to that of image processing but temporal properties of video segments are also studied such as object movement, events, actions and camera motion Abstraction extracts a subset of video data from the original video to facilitate video representation and browsing such as the summary of a video or key frames as the representative of shots

The outputs produced by the above three processes allow for different choices of representations and variations by different people for different purposes This poses a different kind of problem for sharing and interoperability of multimedia content For example, different processes may use different names for the same kinds of outputs, or use the same names for different kinds

Major attention has been directed toward standardising the representation of multimedia content by the MPEG group, which is known as the MPEG-7 standard The representations are compatible with widely divergent multimedia applications In Chapter 2, we detailed the Multimedia Description Schemes and Visual Description for representing the structural, visual and textual features of videos Two different ways of accessing to MPEG-7 have been discussed the text-based and image-based (content-based) solutions The text-based solution studies the statistical properties of terms and the hierarchical structure of documents The image-based solutions rely on the characterisation and pattern matching of low-level visual features

In this thesis, we explore the combined use of both text-based and content-based techniques for MPEG-7 searching An MPEG-7 description is equivalent to a XML document We continued to review three different methods for searching XML documents in Chapter 3, namely the Information Retrieval based, Path Expression and Tree Matching approach The IR-based approach incorporates the properties of document structure into the weighting of document content and relevant information can be found based on full text search alone The latter two approaches focus on the properties of document structures and in specifying document structures in a query as necessary, which turns out to be a disadvantage when the document structures are unknown to users

8.1 Summary of the Aggregated Feature Retrieval Method for MPEG-7

We chose one of the IR-based approaches, namely the aggregation-based method, as a foundation for MPEG-7 retrieval in this work since structural information is incorporated into the weighting of content of document elements and the matching of visual features can be easily adapted to the model

The aggregation-based approach assumes that document structural information is additional evidence to the original content of XML elements It shows us how to combine the content scores of an XML element in its own context and from its structurally related elements using the Ordered Weighting Averaging (OWA) operator and linguistic quantifiers

Following a similar idea, we described an approach in Chapter 4 to index MPEG-7 descriptions for video retrieval based on the assumption that the visual features of shots are seen as the auxiliary evidence that enhance the shots' original meanings We propose to map a video retrieval process to a text retrieval process based on the TF*IDF vector space model via clustering of visual features The indexing process for the visual features is summarised in the following three steps

1 A K-means shot clustering algorithm is used to organise shots of each video based on the content-based / concept-based features

2 We assign a term weight vector to each cluster using a modified TF*IDF algorithm The indexed terms are the spoken words from the ASR transcript in the member shots for the cluster

3 We approximate a new text description of each shot based on its cluster term weight vector and the corresponding cluster-shot distance

Having obtained the newly derived text descriptions for shots via clustering, we aggregate this with the original term weight vector obtained from MPEG-7 textual description for each shot to create its final term weight vector using the Ordered Weighted Averaging operators and linguistic quantifiers

Our video retrieval method for MPEG-7 attempts to handle both text and non-text (i e image-based or concept-based vector) queries We propose to utilise the outputs from the K-means shot clustering algorithm Low-level features are calculated for an image-based query The final query is prepared in three steps

1 Find the N most similar clusters to a given image example based on the low-level features, or to a concept-based query vector based on the concept features
2 Aggregate the term vectors of the N chosen clusters to form a single query term vector
3 Aggregate the original text query and the derived text query (from the non-text query) to create the final query term vector

Retrieval can be done in a straightforward way by calculating the dot products between the aggregated query term vector and the aggregated index vector for each shot and sorting the dot products in descending order

The advantage of the proposed method is the inclusion of semantic meaning (i e terms) from shots that are similar visually within a video to a shot These shots are not necessarily neighbours to the shot They could be from the beginning, middle or end of a video

Experiments were performed as the manual search task on both TRECVID2002 and TRECVID2003 collections and detailed description of the results are given in Chapter 6 and 7, respectively Although some other captured video content such as TV series "The Simpsons" could have been used for evaluation [Browne & Smeaton, 2004], TRECVID video search collections are the only public available collections and we can therefore easily compare our results against other research groups' The

goal of TRECVID is to provide common bases for video retrieval system comparison Development and test video collections for TRECVID experiments used are chosen and made publicly available and query topics are carefully designed to cover a wide range of requests Standardised evaluation methods are used to compare the effectiveness and efficiency of different video retrieval systems

A baseline system using ASR transcripts alone was built for each collection based on the TF*IDF model and ASR transcript that belongs to the shot was used solely in contributing to forming a term weight vector Three types of features were considered in both collections (1) spoken text, (2) concept-based and (3) content-based features

Eight systems with different settings were built for the TRECVID2002 collection and six systems for the TRECVID2003 collection in order to test four evaluation objectives

- Is an aggregated index useful in helping traditional text-only queries?
- Is an aggregated query useful in searching an aggregated index?
- Is there any performance difference between using one best image example and using all image examples in an aggregated query when considering the primitive visual features alone? It would be useful to accept all the example images in a query to avoid the trouble of query selection when users have no knowledge of the search collection
- Is there any performance difference between systems using content-based features alone and systems using concept-based features alone

Firstly, our results show that an aggregated index yields marginally improved results over the baseline using the TRECVID2002 collection, validating our assumption about the visual and semantic similarity of shots in the same cluster The weak semantic similarity existing within a cluster can be used to enrich the original meanings of its constituent shots

On the other hand, our TRECVID2003 results show that there is no significant improvement on employing an aggregated index in video retrieval These outcomes give the opposite position to the TRECVID2002 experiments in which marginal performance improvement was found The main reason for the discrepancy might be different genres of the two search collections being studied

The TRECVID2003 collection mostly contains videos of CNN and ABC broadcast TV news Except for anchorperson, commercial and sports shots, there are few shots remaining that are similar visually and semantically in each news programme Clustering shots within a news programme is thus ineffective compared to clustering within a programme from the TRECVID2002 collection whose videos are documentaries on a specific theme from the 1940s to 1960s and a good amount of shots within each video programme will share similar features

Secondly, both TRECVID 2002 and TRECVID2003 results show that the aggregated query method (i e mapping a content-based query onto a text description and combining it with the original text query) for video retrieval is topic-specific, provided that the image example are from within the search collection If no correct clusters can be found, the derived query vector will be poorly formed In such a case, aggregation was still performed and no pruning technique was used The concept-based queries formed based on the TRECVID concepts were too general to find any appropriate or relevant *topK* similar clusters for query mapping

Thirdly, both TRECVID 2002 and TRECVID2003 results show that using multiple image examples in a query is not as effective as using one single best image example It is seen that using multiple images in a query for video retrieval introduces more unrelated query terms and the additional query terms appear to cause a slight reduction in precision

Finally, additional experiments were developed to evaluate the performance difference when using only content-based features and only concept-based features in a query for video retrieval Our results show that using a derived text query from primitive visual features of an image query (without including the original text

query) is feasible in video retrieval But the mapping of a concept-based query to a text query description is less feasible due to the generality of the defined concepts we evaluated and duplicate formulation of different queries

In summary, our approach is shown to be useful when each video in a collection contains a high proportion of shots that have visual and semantic similarity such as the documentaries in the TRECVID2002 collection Adding meaning to a shot based on the shots that are around it might be an effective method for video retrieval when each video has low proportion of similar shots (i e TV news programmes due to the fact that neighbouring shots typically fall within the same story)

8.2 Problems of the Method

There are two main problems in our proposed retrieval approach for MPEG-7 The first problem relates to the nature of clustering It is understood that different clustering processes create different partitions In order to summarise the meaning of a cluster, we simply collect spoken terms from each shot within the cluster and calculate the weights for the terms

Our experiments have shown that the mapping process of visual features of a shot to a text vector via clustering during the indexing stage is problematic Certain clusters may be well formed and shots in such a cluster present visual and semantic similarity Meanwhile, totally unexpected clusters might be created in which shots present little similarity and shots in these groupings will not benefit from the addition of cluster terms, but would create misleading information

Another problem arises when we attempt to map a non-text query onto a query term vector in the query preparation stage Such a mapping process allows the non-text query to expand into either a good textual description with meanings closely related to a user's information need, or into a poor description with meanings completely different from the need

The main difficulty here is to find the most similar clusters to the given non-text query and this is mostly affected by the quality of query images If a given image query is from within the clusters, the mapped text query will give performance benefit

Another difficulty with the query mapping process is the inclusion of more query terms, some of which are related and some are unrelated to the need In such a case, aggregation with the original text query was still performed and no pruning technique to filter the unrelated terms in the description was used In order to reduce the influence of derived query terms which are unrelated to the topic, more weight is assigned to the original query terms prior to the query aggregation process The main meanings of the final aggregated query will thus remain so that its system performance can be kept as good as one using the original query terms only

8.3 Future Work

Much research in digital video retrieval attempts to mimic and understand video content through automated equipment Standards such as the MPEG family are developed for this very purpose These are important interim steps to smooth the transition from the analogue world to on-line digital video services for searching, sharing and interoperability In order to build an operational video retrieval system, related storage and retrieval methods have to take into account the different features of videos that are meaningful and detectable For many years researchers have produced some fruitful results with the tasks of video parsing, feature extraction and abstraction We are now facing the challenge of finding an approach for video retrieval with reasonably good performance using these detected features

We have seen the existing semantic gap between the text and low-level features which impedes the progress and achievement of video retrieval Researchers increasingly talk about ontological concepts as newly defining features of video content and much attention is devoted to the development of such concept detectors

The goal of the concept detection is to discover the degree to which the similarity of content-based feature measures implies the inheritability of concepts

A choice of concepts is the first step in designing things that can be represented in video shots as well as in users' queries Any incompleteness or restrictions in the concepts inevitably limit the generality of videos that uses those concepts (i e all genres of video programmes, including news, documentaries, drama and comedy) Limited concepts will be useful for single genres in highly specialised domains such as finding the "anchorpersons", "sporting event" and "weather news" in news programmes But to share concept representations with other genres, the selected concepts must be embedded within a more general framework top-level concepts form the hierarchy that can relate the details of the lower-level ones For example, in the TRECVID feature detection task, the "news subject face" concept can be seen as a sub-concept of the "people" concept, and the "building" concept appears as part of the "outdoor" concept

A concept-based feature is often seen as fast representation of video shots and shots that pass through the concept can be further examined either by users or by using the low-level feature representations during a search If a user's chosen concept is too general such as "people", the number of filtered shots for further examination remains over half of the video collection simply because most video shots contain one or more persons If the concept is too specific such as "Madeleine Albright", it would be a reasonably distinguishable feature for shot representation and would be useful to answer very specific queries only

The design of concept features not only depends on the generality level that most users require for formulating their queries, but more importantly on the difficulty of the automatic extraction process The accuracy of the concept detection output varies based on the number of low-level feature patterns used to summarise a concept during a training (i e learning) process For example, the precision of detecting the "weather news" and "anchorpersons" shots is high in TRECVID experiments due to the countable and predictable visual patterns of both concepts occurring in news programmes On the contrary, the precision of extracting the

"building" and "animal" shots is relatively low in that the number of possible visual patterns for representing the two concepts is large and unpredictable A building could be either tall with glass wall or small with concrete wall, an animal could be close-ups of single animals or a further view of a group

In this thesis we described our proposed video retrieval approach for MPEG-7 descriptions by using both text and visual features extracted from video content We made some simple and explicit assumptions about the semantic and visual similarity of shots in each video The approach overall is not a straightforward one but consists of various processes (i e shot clustering and term assignment for visual features), each of which introduces desired as well as unwanted information into the final shot representations Our TRECVID experiments have shown that the approach gives marginally improvement over the baseline when each video in a collection contains a high proportion of shots that have visual and semantic similarity such as documentaries

If we are to succeed in delivering maximum performance from the conceptual feature detection for video content in the future, we believe that our approach using the concept-based features for clustering shots and creating an aggregated index could lead to acceptable performance for video retrieval This is because detecting concept-based features is the first step to bridging the semantic gap between text and low-level features The shot clustering output is subject to the availability and coverage of the defined concepts to a search collection being indexed Shots could be put together simply because of the absence of most concepts in the shots and therefore no conceptual similarity is presented in the created cluster

Furthermore, the next step in future work is to develop a better shot clustering algorithm so as to reduce the number of unexpected clusters (i e shots in the cluster shows no visual and semantic similarity) being created We also wish to explore if there is a pruning technique for filtering out terms that has little or no use for summarising cluster meanings and approximating extra shot meanings during the

creation of an aggregated index How to handle the pruning process intelligently is not clear, but is a very real problem

The query preparation process is another component of our overall video retrieval system Using a derived text query from a concept-based query in video retrieval is likely to be poor because of the difficulty of finding the *topK* most similar clusters using concepts The concepts are so general that clusters can not be distinguished Furthermore, the construction of a concept-based query using 3 scales {0, 0 5, 1} allows the probability of formulating the same textual query for different topics

On the other hand, our experiments have shown that using a text query derived from a content-based query is possible, provided that the image examples are from within the same search collection We have proposed a way of deriving a text query based on the *topK* most similar clusters to an image example, and hope to implement a pruning technique to remove the derived query terms irrelevant to information need

A direction for future work is to use the strength of both content-based and concept-based features in different processes For the indexing preparation, we would use concept-based features for shot clustering to obtain groupings in which shots could present reasonable semantic similarity, and the aggregated index would then be created in the same way as described in this thesis A step would be introduced to calculate another set of cluster centroids based on the content-based features of shots in the created clusters The resulting centroids would be used to assist the query preparation process Given an image example, the corresponding low-level features would be extracted and used to search the set of cluster centroids obtained based on the content-based features The remaining steps of deriving a text query based on the *topK* most similar clusters would be the same as described in this thesis In such a way, we hope to better understand the usage of both types of features in video retrieval

Finally, the use of conceptual features in video retrieval has been the source of a great deal of controversy To some, it represents the beginning of greater interdisciplinary research that could provide ready access to the detection techniques, promoting the role of conceptual features in video retrieval, while to others, it is the challenges in the development that led the idea to become more critical We would like to examine the retrieval performance of our approach against different level of accuracy and coverage of the extracted concepts to see whether better performance can be gained when the concepts are accurately detected Currently, there is no existing test collection which facilitates this If we are to succeed in developing such a collection, we would be able to better understand the position of concept-based features in video retrieval

BIBLIOGRAPHY

[Adams et al, 2002] Adams, B , Amir, A , Dorai, C , Ghosal, S , Lyengar, G , Jaimes, A , Lang, C , Lin, C , Natsev, A , Naphade, M , Neti, C , Nock, H J , Permuter, H H , Singh, R , Smith, J R , Srinivasan, S , Tseng, B L , Ashwin,T V and Zhang, D IBM Research TREC-2002 Video Retrieval System The 11th Text Retrieval Conference, Gaithersburg, November 2002 Available at http //trec nist gov/pubs/trec11/papers/ibm smith vid pdf, last visit on 8 May 2004

[Anderberg, 1973] Anderberg, M R Cluster Analysis for Applications Academic Press, London, 1973

[Amir et al, 2003] Amir, A , Hsu, W , Iyengar, G , Lin, C -Y , Naphade, M , Natsev, A , Neti, C , Nock, H J , Smith, J R , Tseng, B L , Wu, Y , Zhang, D IBM Research TRECVID-2003 System In Proceedings of NIST Text Retrieval Conference (TREC), Gaithersburg, MD, November 2003 Available at http //www-nlpir nist gov/projects/tvpubs/papers/ibm smith paper final2 pdf, last visit on 8 May 2004

[Barras et al, 2002] Barras, C , Allauzen, A , Lamel, L and Gauvain, J L Transcribing Audio-Video Archives In the Proceedings of ICASSP, pages 13-16, Orlando, May 2002

[Bober, 2001] Bober, M MPEG-7 Visual Shape Descriptors In IEEE Transactions on Circuits and Systems for Video Technology, Vol 11, No 6, June 2001, 716-719

[Bolle et al, 1998] Bolle, R M , Yeo, B -L and Yeung, M M Video Query Research Directions In IBM Journal of Research and Development, Multimedia Systems, Vol 42, No 2, 1998 Available at http //www research ibm com/journal/rd/422/bolle html, last visit on 8 May 2004

[Browne et al, 2000] Browne P , Smeaton, A , Murphy, N , O'Connor, N , Marlow, S and Berrut, C Evaluating and Combining Digital Video Shot Boundary Detection Algorithms IMVIP 2000 - Irish Machine Vision and Image Processing Conference, Belfast, Northern Ireland, 2000 Available at http //www cdvp dcu ie/Papers/IMVIP2000 pdf, last visit on 8 May 2004

[Browne & Smeaton, 2004] Browne P & Smeaton, A Video Information Retrieval Using Objects and Ostensive Relevance Feedback SAC 2004 – ACM Symposium on Applied Computing, Nicosia, Cyprus, 14-17 March 2004

[Calic et al, 2002] Calic, J , Sav, S , Izquierdo, E , Marlow, S , Murphy, N and O'Connor, N Temporal Video Segmentation for Real-Time Key Frame Extraction ICASSP 2002 - International Conference on Acoustics, Speech and Signal Processing, Orlando, Florida, 13-17 May 2002 Available at http //www cdvp dcu ie/Papers/ICASSP2002 pdf, last visit on 8 May 2004

[Chen & Chua, 2001] Chen, L and Chua, T S A match and tiling approach to content-based image retrieval In Proceeding of ICME, 2001

[Chen et al, 2001] Chen, Z , Jagadish, H V , Korn, F , Koudas, N , Muthukrishnan, S , Raymond Ng, and Srivastava, D Counting twig matches in a tree In Proceedings of IEEE International Conference on Data Engineering, Heidelberg, Germany, April 2001, 595-604

[Chiariglione, 1996] Chiariglione, L Short MPEG-1 Description, June 1996 Available at http //www chiariglione org/mpeg/standards/mpeg-1/mpeg-1 htm, last visit on 8 May 2004

[Chiariglione, 2000] Chiariglione, L Short MPEG-2 Description, Oct 2000 Available at http //www chiariglione org/mpeg/standards/mpeg-2/mpeg-2 htm, last visit on 8 May 2004

[Christel et al, 1997] Christel, M , Winkler, D , and Taylor, C Multimedia Abstractions for a Digital Video Library In Proceedings of the 2nd ACM International Conference on Digital Libraries, Philadelphia, PA, July 1997, 21-29

[Cieplinski et al, 2001] Cieplinski, L , Kim, M , Ohm, J R , Pickering, M and Yamada, A ISO/IEC 15938-3/FCD Information Technology – Multimedia Content Description Interface – Part 3 Visual, working document N4062, Singapore March, 2001

[Cleverdon et al, 1966] Cleverdon, C W , Mills, J and Keen, E M Factors Determinging the Performance of Indexing Systems, vol 1 - design, Aslib Crandfield Research Project, Crandfield, England, 1966 Available at http //www itl nist gov/iaui/894 02/projects/irlib/pubs/cranv1p1/cranv1p1_index/cranv1p1 html, last visit on 8 May 2004

[Czirjek et al, 2003] Czirjek, C , O'Connor, N , Marlow, S , Murphy, N Face Detection and Clustering for Video Indexing Applications Acivs 2003 - Advanced Concepts for Intelligent Vision Systems, Ghent, Belgium, 2-5 September 2003

[Day & Martinez, 2002] Day, N and Martinez, J M (ed) Introduction to MPEG-7 (v4 0), working document N4675, (2002) Available at http //mpeg telecomitalialab com/working_documents htm, last visit on 29th Oct, 2002

[Deerwester et al, 1990] Deerwester, S , Dumais, S T , Furnas, G W , Landauer, T K , & Harshman, R Indexing by latent semantic analysis Journal of the American Society for Information Science, 41(6), 1990, 391-407

[Everitt, 1980] Everitt, B Cluster Analysis, 2nd edition Halsted Press, New York, 1980

[Faloutsos, et al, 1994] Faloutsos, C , Barber, R Flickner, M , Niblack, W , Petkovic, D and Equitz, W Efficient and effective querying by image content Journal of Intelligent Information Systems, 3(3/4), July 1994, 231-- 262

[Fuhr & Rolleke, 1998] Fuhr, N and Rolleke, T HySpirit - a Flexible System for Investigating Probabilistic Reasoning in Multimedia Information Retrieval In Proceedings of the 6th International Conference on Extending Database Technology (EDBT), Valencia, Spain, 1998, 24-38

[Gauvain et al, 2000] Gauvain, J L , Lamel, L , Barras, C , Adda, G , and Kercardio, Y The LIMSI SDR System for TREC-9 In the 9th Text Retrieval Conference, Gaithersburg, Nov 2000 Available at http //trec nist gov/pubs/trec9/papers/limsi-sdr00 pdf, last visit on 8 May 2004

[Gauvain et al, 2002] Gauvain, J L , Lamel, L , and Adda, G The LIMSI Broadcast News Transcription System In Speech Communication, 37(1-2), 2002, 89-108 Available at ftp //tlp limsi fr/public/spcH4_limsi ps Z, last visit on 8 May 2004

[Graves & Lalmas, 2002] Graves, A and Lalmas, M Video Retrieval using an MPEG-7 Based Inference Network In Proceedings of SIGIR'02, Tampere, Finland, August 2002, 339-346

[Guttman, 1984] Guttman, A R-Trees A Dynamic Index Structure for Spatial Searching In Proceedings of ACM SIGMOD International Conference on Management of Data, Boston, MA, 1984, 47-57

[Harman, 1992] Harman, D Overview of the First Text Retrieval Conference In the Proceedings of the First Text REtrieval Conference (TREC-1), Gaithersburg, Maryland, November 1992 Available at http //trec nist gov/pubs/trec1/papers/01 txt, last visit on 8 May 2004

[Hauptmann et al, 2003] Hauptmann, A , Baron, R V , Chen, M -Y , Christel, M , Duygulu, P , Huang, C , Jin, R , Lin, W -H , Ng, T , Moraveji, N , Papernick, N , Snoek, C G M , Tzanetakis, G , Yang, J , Yang, R and Wactlar, H D Informedia at TRECVID 2003 Analyzing and Searching Broadcast News Video In The 12th Text Retrieval Conference, Gaithersburg, November 2003 Available at http //www-nlpir nist gov/projects/tvpubs/papers/cmu final paper pdf, last visit on 8 May 2004

[Heesch et al, 2003] Heesch, D , Pickering, M J , Ruger, S , and Yavlinsky, A Video Retrieval within a Browsing Framework using Key Frames In The 12th Text Retrieval Conference, Gaithersburg, November 2003 Available at http //www doc ic ac uk/~mjp3/phd/www-pub/trec2003-nb pdf, last visit on 8 May 2004

[Jain & Dubes, 1988] Jain, A K and Dubes, R C Algorithms for Clustering Data Prentice Hall, New Jersey, 1988

[Jain et al, 1999] Jain, A K , Murty, M N and Flynn, P J Data Clustering A Review ACM Computing Survey, Vol 31, No 3, September 1999

[Kazai et al, 2001a] Kazai, G , Lalmas, M and Roelleke T A Model for the Representation and Focussed Retrieval of Structured Documents based on Fuzzy Aggregation In String Processing and Information retrieval (SPIRE 2001) Conference, Laguna De San Rafael, Chile, November 2001

[Kazai et al, 2001b] Kazai, G , Lalmas, M and Roelleke T Aggregated Representation for the Focussed Retrieval of Structured Documents In SIGIR'01 Workshop, Mathematical/Formal Methods in IR, 2001

[Koenen, 2002] Koenen, R Overview of the MPEG-4 Standard, WG11\N4668, March 2002 Available at http //www chiariglione org/mpeg/standards/mpeg-4/mpeg-4 htm, last visit on 25 Oct, 2002

[Lalmas et al, 2001] Lalmas, M , Mory, B , Moutogianm, E , Putz, W and Roelleke T Searching Multimedia Data Using MPEG-7 Descriptions in a Broadcast Terminal 2001 Available at http //citeseer nj nec com/449117 html, last visit on 8 May 2004

[Lienhart et al, 1998] Lienhart, R , Effelsberg, W and Jam, R VisualGrep A systematic method to compare and retrieve video sequences In Proc SPIE Storage and Retrieval for Image and Video Databases, vol 3312, 1998, 271-282

[Liu et al, 1999] Liu X , Zhuang Y & Pan Y A New Approach to Retrieve Video by Example Video Clip In Proceeding of the 7th ACM International Multimedia Conference (Multimedia 99) Poster Session Orlando, Florida, USA October 30-November 5, 1999

[Luhn, 1958] Luhn, H P The Automatic Creation of Literature Abstracts IBM Journal of Research and Development, Vol 2, No 2, April 1958, 159-165 Available at http //domino research ibm com/tchjr/journalindex nsf/0/97e0420635a5000a85256 bfa00683d33?OpenDocument, last visit on 8 May 2004

[Lyman & Varian, 2003] Lyman, P and Varian, Hal R How Much Information? 2003 Available at http //www sims berkeley edu/research/projects/how-much-info-2003/, last visit on 8 May 2004

[Manjunath et al, 2001] Manjunath, B S , Ohm Jens-Rainer, Vasudevan, V V and Yamada, A Color and Texture Descriptors In IEEE Transactions on Circuits and Systems for Video Technology, Vol 11, No 6, June 2001, 703-715

[Marchionim, 1995] Marchionini, G Information Seeking in Electronic Environments, Cambridge University Press, 1995

[Martinez, 2003] Martinez, J M MPEG-7 Overview, ISO/IEC JTC1/SC29/WG11N5525, Pattaya, March 2003 Available at http //www chiariglione org/mpeg/standards/mpeg-7/mpeg-7 htm, last visit on 8 May 2004

[McHugh et al, 1997] McHugh, J , Abiteboul, S , Goldman, R , Quass, D and Widom, J Lore a Database Management System for Semistructured Data Technical Report, Stanford University, Database Group, Feb 1997 Available at http //citeseer nj nec com/mchugh97lore html, last visit on 8 May 2004

[McHugh et al, 1998] McHugh, J , Widom, J , Abiteboul, S , Luo, Q and Rajaraman, A Indexing Semistructured Data Technical Report, Stanford University, Computer Science Department, January 1998 Available at http //citeseer nj nec com/mchugh98indexing html, last visit on 8 May 2004

[MPEG7-N2466] Licensing Agreement for the MPEG-7 Content Set, MPEG98/N2466, Atlantic City, USA, 1998 Available at http //www tnt uni-hannover de/project/mpeg/audio/public/mpeg7/w2466 pdf, last visit on 8 May 2004

[MPEG7-N4039] MPEG-7 Interoperability, Conformance Testing and Profiling (Version 2 0), working document N4039, March 2001 Available at http //www chiarighone org/mpeg/working_documents/mpeg-07/general/interop zip, last visit on 5th Feb 2004

[Ngo et al, 2001] Ngo, C W , Pong, T C and Zhang, H J On Clustering and Retrieval of Video Shots ACM Multimedia, 2001, 51-60

[Pearmain et al, 2002] Pearmain, A , Lalmas, M , Moutogianni, E , Papworth, D , Healy, P and Roelleke T Using MPEG-7 at the Consumer Terminal in Broadcasting In EURASIP Journal on Applied Signal Processing, Vol 2002, No 4, April 2002, 354-361

[Pickering et al, 2002] Pickering, M J , Heesch, D , O'Callaghan, R , Ruger, S , and Bull,D Video Retrieval Using Global Features in Keyframes In The 11th Text Retrieval Conference, Gaithersburg, 5 November, 2002 Available at http //trec nist gov/pubs/trec11/papers/imperial pickering pdf, last visit on 8 May 2004

[Press et al, 1989] Press, W H , Flannery, B P , Teukolsky, S A , and Vetterling, W T Fourier Transform of Discretely Sampled Data §12 1 in Numerical Recipes in C The Art of Scientific Computing Cambridge, England Cambridge University Press, 1989, 500-504

[Rautiainen et al, 2002] Rautiainen, M , Penttila, J , Vorobiev, D , Noponen, K , Vayrynen, P , Hosio, M , Matinmikko, E , Makela, S , Peltola, J , Ojala, T and Seppanen, T TREC 2002 Video Track Experiments at MediaTeam Oulu and

VTT In The 11th Text Retrieval Conference, Gaithersburg, 5 November, 2002 Available at http //trec nist gov/pubs/trec11/papers/uoulu rautiainen pdf, last visit on 8 May 2004

[Rautiamen et al, 2003] Rautiainen, M , Noponen, K , Hosio, M , Koskela, T , Liu, J , Ojala, T , Seppanen, T , Penttila, J , Piertarila, P , Makela, S M and Peltola, J TRECVID 2003 Experiments at Media Team Oulu and VTT In The 12th Text Retrieval Conference, Gaithersburg, Nvember 2003 Available at http //www-nlpir nist gov/projects/tvpubs/papers/uoulu paper pdf, last visit on 8 May 2004

[Rijsbergen, 1979] Van Rijsbergen, C J Information Retrieval, 2nd edition, Butterworths, London, 1979

[Rizzolo & Mendelzon, 2001]Rizzolo, F and Mendelzon, A Indexing XML Data with ToXin In Proceedings of Fourth International Workshop on the Web and Databases, 2001 Available at http //citeseer nj nec com/rizzolo01indexing html, last visit on 8 May 2004

[Sadlier et al, 2003] Sadlier, D , O'Connor, N , Marlow, S , and Murphy, N A Combined Audio-Visual Contribution to Event Detection in Field Sports Broadcast Video Case Study Gaelic Football ISSPIT'03 - IEEE International Symposium on Signal Processing and Information Technology, Darmstadt, Germany, 14-17 December 2003

[Salembier & Smith, 2001] Salembier, P and Smith, J R MPEG-7 Multimedia Description Schemes In IEEE Transactions on Circuits and Systems for Video Technology, Vol 11, No 6, June 2001, 748-759

[Salton & McGill, 1983] Salton, G and McGill, M Introduction to Modern Information Retrieval McGraw-Hill, 1983

[Saracevic, 1975] Saracevic, T Relevance A Review of and a Framework for the Thinking on the Notion in Information Science In Journal of the American Society for Information Science, 26(6), 1975, 321-343 Available at http //www scils rutgers edu/~tefko/Saracevic_relevance_75 pdf, last visit on 8 May 2004

[Saracevic, 1996] Saracevic, T Relevance Reconsidered In Information Science Integration in Perspectives, Proceedings of the Second Conference on Conceptions of Library and Information Science Copenhagen (Denmark), 1996, 201-218 Available at http //www scils rutgers edu/~tefko/CoLIS2_1996 doc, last visit on 8 May 2004

[Sato et al, 1998] Sato, T , Kanade, T , Hughes, E , and Smith, M Video OCR for Digital News Archive In Proceeding Workshop on Content-Based Access of Image and Video Databases, Los Alamitos, CA, Jan 1998, 52-60 Available at http //www informedia cs cmu edu/documents/vocr_acm98 pdf, last visit on 31 August 2004

[Schlieder & Naumann, 2000] Schlieder, T and Naumann, F Approximate tree embedding for querying XML data In SIGIR'00 Workshop on XML and Information Retrieval, Athens, Greece, July 2000

[Sheridan & Smeaton, 1992] Sheridan, P and Smeaton, A F The Application of Morpho-Syntactic Language Processing to Effective Phrase Matching Information Processing and Management, 28(3), 1992, 349-370

[Sikora, 2001] Sikora, T The MPEG-7 Visual Standard for Content Description-an Overview In IEEE Transactions on Circuits and Systems for Video Technology, Vol 11, No 6, June 2001, 696-702

[Smeaton et al, 2001] Smeaton, A F , Murphy, N , O'Connor, N , Marlow, S , Lee, H , Mc Donald, K , Browne, P and Ye, J The Fischlar Digital Video System A Digital Library of Broadcast TV Programmes JCDL 2001 - ACM+IEEE Joint Conference on Digital Libraries, Roanoke, VA, 24-28 June 2001

[Smeaton & Over, 2002] Smeaton, A F and Over, P The TREC 2002 Video Track Report In The 11th Text Retrieval Conference, Gaithersburg, 5 November 2002 Available at http //www-nlpir nist gov/projects/t2002v/results/notebook papers/VIDEO OVERVIEW pdf, last visit on 8 May 2004

[Smeaton et al, 2003] Smeaton, A F , Kraaij, W , and Over, P TRECVID 2003 - An Introduction TRECVID 2003 - Text REtrieval Conference TRECVID Workshop, Gaithersburg, Maryland, 17-18 November 2003 Available at http //www-nlpir nist gov/projects/tvpubs/papers/tv3intro paper pdf, last visit on 8 May 2004

[Stokes et al, 2004] Stokes, N , Carthy, J , and Smeaton, A SeLeCT A Lexical Cohesion based News Story Segmentation System In Journal of AI Communications, Vol 17, No 1, 2004, 3-12

[Tai, 1979] Tai, K -C The Tree-to-Tree correction problem Journal of the ACM, 26(3), 1979, 422-433

[Turtle & Croft, 1991] Turtle, H and Croft, W Evaluation of an inference network-based retrieval model In ACM Transactions on Information Systems 9, 1991, 187-222

[Voorhees, 2002] Voorhees, E M Overview of TREC 2002 In the Proceedings of the 11th Text REtrieval Conference (TREC-1), Gaithersburg, Maryland, November, 2002 Available at http //trec nist gov/pubs/trec11/papers/OVERVIEW 11 pdf, last visit on 8 May 2004

[Westerveld et al, 2002] Westerveld, T , Vries, A P , Ballegooij, A CWI at the TREC-2002 Video Track In the 11th Text Retrieval Conference, Gaithersburg, 5 November, 2002 Available at http //trec nist gov/pubs/trec11/papers/cwi westerveld pdf, last visit on 8 May 2004

[Westerveld et al, 2003] Westerveld, T , Vries, A P , Ballegooij, A , Jong, F and Hiemstra, D A Probabilistic Multimedia Retrieval Model and Its Evaluation In EURASIP Journal on Applied Signal Processing *2003 2*, Special Issue on Unstructured Information Management from Multimedia Data Sources, ISSN 1110-8657, Hindawi, 2003

[Westerveld et al, 2003] Westerveld, T , Vries, A P , Ianeva, T , Boldareva, L and Hiemstra, D Combining Information Sources for Video Retrieval In the 12th Text Retrieval Conference, Gaithersburg, November 2003 Available at http //www-nlpir nist gov/projects/tvpubs/papers/lowlandsteam final2 paper pdf, last visit on 8 May 2004

[Wilkinson, 1994] Wilkinson, R Effective retrieval of structured documents In Proceedings of SIGIR'94, Dublin, Ireland, 1994

[Wolf et al, 2002] Wolf, C , Doermann, D and Rautiainen, M Video Indexing and Retrieval at UMD In The 11th Text Retrieval Conference, Gaithersburg, 5 November, 2002 Available at http //trec nist gov/pubs/trec11/papers/umd doermann pdf, last visit 8 May 2004

[Yager, 2000] Yager, R A Hierarchical Document Retrieval Language Information Retrieval, Vol 3, No 4, December 2000, 357-377

[Ye & Smeaton, 2003] Ye, J and Smeaton, A F Aggregated Feature Retrieval for MPEG-7 Poster in the Proceedings of 25th European Conference on IR Research (ECIR), Pisa, Italy, April 2003, 563-570

[Yeung et al, 1996] Yeung, M M , Yeo, B , and Liu, B Extracting story units from long programs for video browsing and navigation In Proceedings of IEEE Multimedia Computing & Systems, 1996, 296--305

[Zhang et al, 1995] Zhang, H , Low, C , Smoliar, S and Wu, J Video parsing, retrieval and browsing an integrated and content-based solution In Proceedings of 3rd ACM International Conference on Multimedia (MM '95), San Francisco, CA, 5-9 November, 503-512, 1995

[Zobel, 1998] Zobel, J How Reliable are the Results of Large-Scale Information Retrieval Experiments? Croft, W B , Moffat, A , Rijsbergen, C J , Wilkinson, R and Zobel, J , editors, Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 307-314, Melbourne, Australia, August 1998 ACM Press, New York