COMBINING MULTIMODAL EXTERNAL RESOURCES FOR EVENT-BASED NEWS VIDEO RETRIEVAL AND QUESTION ANSWERING

SHI-YONG NEO

(B. COMP (HONORS), NATIONAL UNIVERSITY OF SINGAPORE)

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE

SCHOOL OF COMPUTING

NATIONAL UNIVERSITY SINGAPORE

2008

Dedication

To Wendy and Cheran

Acknowledgements

First, I would like to thank my supervisor Tat-Seng Chua, for his great guidance over the last six years. Thinking back, I was just an average undergraduate student when he gave me the invaluable opportunity to join the PRIS group as an undergraduate student researcher in 2002. I was deeply inspired by his love and commitment towards the field of multimedia research. What I learned from him is not just techniques in multimedia content analysis, but more importantly, self development, time management and communication skills that will benefit me for life. I also appreciate the freedom I was given to work with different collaborators in NUS and ICT (China), which has greatly broadened my understanding across other research areas.

I would also like to thank my other thesis committee members, Mohan Kankanhalli, Wee-Kheng Leow and Ye Wang, for their invaluable assistance, feedback and patience at all stages of this thesis. Their criticisms, comments, and advice were critical in making this thesis more accurate, more complete and clearer to read. I am also grateful to the financial support given by SMF (Singapore Millennium Foundation) and Temasek Holdings.

Moreover, I am also indebted to fellow group members in NUS for providing me inspiration and suggestions during the meetings. My special thanks go to Hai-Kiat Goh, Yan-Tao Zheng, Huanbo Luan, Renxu Sun and Xiaoming Zhang for their insightful discussions. Their great guidance helped me tremendously in understanding the area of multimedia information retrieval.

Last, but definitely not the least, I would also like to thank my family especially my wife Wendy, for their love and support.

Contents

Summary

The ever-increasing amount of multimedia data available online creates an urgent need on how to index these information sources and support effective retrieval by users. In recent years, we observe the gradual shift from performing retrieval solely based on analyzing one media source at a time, to fusion of diverse knowledge sources from correlated media types, context and language resources. In particular, the use of Web knowledge has increased, as recent research has shown that the judicious use of such resources can effectively complement the limited extractable semantics from the video source alone. The new challenge faced by the multimedia community is therefore how to *obtain* and *combine* such diverse multimedia knowledge sources. While considerable effort has been spend on extracting valuable semantics from targeted multimedia data, less attention has been given to the problem of utilizing external resources around such data and finding an effective strategy to fuse them. In addition, it is also essential to develop principled fusion approaches that can leverage query, content and context information automatically to support precise retrieval.

This thesis presents how we leverage external knowledge from the Web to complement the extractable features from video contents. In particular, we develop an event-based retrieval model that acts as a principled framework to combine the diverse knowledge sources for news video retrieval. We employ the various online news websites and news blogs to supplement details that are not available in news video and extract innate relationship between different content entities during data clustering.

The event-based retrieval uses query class dependent models which automatically discover fusion parameters for fusing multimodal features based on previous retrieval

results, and predicts parameters for unseen queries. Other external resources like online lexical dictionary (WordNet) and photo sharing site (Flickr) are also used to inference linkages between query terms and semantic concepts in news video. Hierarchical clustering is then carried out to discover the latent structure of news (topic hierarchy). This newly discovered topic hierarchy facilitates effective browsing through key news events and precise question answering.

We evaluate the proposed approaches using the large-scale video collections available from TRECVID. Experimental evaluations demonstrate promising performance as compared to other state-of-the-art systems. In addition, the system is able to answer other related queries in a question-answering setting through the use of the topic hierarchy. User studies indicate that the event-based topic browsing is both effective and appealing. Even though this work is carried out mainly on news videos, many of the proposed techniques such as the event feature representation, query expansion and the use of high-level-features in query processing can also be applied to retrieval of other video genres such as the documentaries and movies.

List of Tables

List of Figures

Notations

s	shot
S	set of all shots $s_j \in S$ arbitrary chosen shot j in S
fs	feature vector of a shot
v	news video story
V	set of all news video stories, $v_j \in V$ arbitrary chosen news video story j in V
fv	feature vector of a news story
a	text article
A	set of all text articles, $a_j \in A$ arbitrary chosen text article j in A
fa	feature vector of text article
D_s	matrix of near duplicate for all shots, size of $\|S\|\times\|S\|$, {1- yes, 0-no}
D_v	matrix of near duplicate for all stories, size of $\|V\|\times\|V\|$, {1- yes, 0-no}
CD	cluster density
CV	cluster volume space
CRT	cluster representative template
TP	cluster partition (time-based)
e	event entities in a cluster template
C	cluster
c	cluster centroid
Q	query
q	query terms
q'	expanded query terms
q_{images}	query images or video key-frames provide by user
HLF_k	a particular high level feature
$conf$	confidence, normalized [0,1]
i,j,k,l,n	arbitrary numbers
α,β	arbitrary parameters
w	arbitrary word

Chapter 1

Introduction

With the ever-increasing amount of multimedia data, effective multimedia information retrieval is becoming increasingly important especially. Such massive amount of multimedia data requires intelligent systems that are capable of retrieving what the users need accurately and in a timely fashion. In a recent study by CacheLogic [Cach07] a network infrastructure company, the current Web is multimedia dominant, as video and audio data transfer accounts for 70% of the total internet traffic for year 2006. Besides, it is also imminent that this percentage will increase, given the fast growing data available from information sharing sites such as YouTube and Google video. [Rowe04] however pointed out that if multimedia data present on the Web is not manageable and accessible by general users, it is highly likely that they will become redundant or unnecessary. It is therefore essential to develop techniques to index multimedia data effectively so that such information can be made retrievable.

Simultaneously, while we ponder over how to improve indexing and retrieval, we are yet to make effective use of external sources of information relating to the data source to supplement the tasks. The vast collections of different multimodal data available on the Web can sometimes provide complementary features or valuable collective knowledge that can facilitate retrieval. One such external feature is the famous PageRank algorithm [Brin98] implemented in Google search. The technique leverages the linking information between

web pages to determine the importance of a web page. Another commonly used knowledge is *popularity*. We can accurately predict, for example, who are the top singers or top songs by looking at the number of uploaded/downloaded songs from a MP3 website. This popularity information, which is not available from the source (i.e. song, video, podcast), can influence and help the general user in searching for what they might want. In addition, the Web also contains abundant information in both text and video for more structured types of information such as the news and sports. Research has shown that the use of external text articles to correct erroneous speech transcripts or closed captions [Wact00, Yang03, Zhao06] from news video sources are effective.

The new problem that our multimedia community faces now is how to *obtain* and *combine* such diverse multimedia knowledge sources. While considerable effort has been expended on extracting valuable semantics from the targeted multimedia data, relatively little attention has been given to the problem of utilizing relevant external resources around such data. There is thus a strong need to shift the paradigm for data analysis from using only one data source, to the fusion of diverse knowledge sources. For example, searching *"a scene of flood"* in a news video collection might leverage information from one of these contexts or their combination: (a) looking for the presence of *"water-bodies"* in the video or frames; (b) identifying the speech segments that mention terms like *"flood, rain, etc…"* (c) utilizing prior knowledge if available, such as the *location* or *dates* of such events (i.e. flood), and (d) searching for news videos that mention these location around the eventful dates. In fact, it is possible to obtain such prior knowledge of locations and dates arising from a certain event with good accuracy from text collection that is available online.

In this thesis, we apply our discovered indexing and retrieval techniques mainly to

the domain of multimedia news video. We will elaborate in detail the issues of *how to obtain* (extracting usable semantics from external data) and *how to combine* (develop effective combination strategies to merge multiple knowledge sources) with respect to the proposed event-based model, followed by summarizing the contributions of this thesis.

1. 1 Leveraging Multi-source External Resources

At present, the limited amount of video semantics obtainable from within news video is not sufficient to support precise retrieval. This is because news video is often presented in a summarized form and various important contexts may not be available. In addition, available features such as the speech transcripts from ASR (automatic speech recognition) may be erroneous. In this work, we propose to supplement news video retrieval with various external resources. Prior works like [Kenn05, Neo06, Volk06] utilized language resources to help relating query to available features. [Chen04, Neo05, Zhao06] relied on parallel news information to supplement features, and more recently [Neo07] utilized collective knowledge for fusion of retrieval with general human interest. In this thesis, we explore four diverse sources of online information and describe how to make use of these resources to supplement retrieval.

Language resource. The use of online language resources such as the lexical dictionary WordNet [Fell98] has shown to be very effective in complementing text retrieval [Trec]. This online lexical reference system whose design is inspired by the current psycholinguistic theories of human lexical memory provides linguistic features such as gloss, word senses, synonyms and hyponyms. Based on this thesaurus, we are able to infer lexical semantic relation from query terms to gather additional context. One such example is

as follows, given 2 sets of words {car, boat} and {water}, where we can utilize their lexical definitions such as "car" is "a motor vehicle with four wheels; usually propelled by an internal combustion engine" and "boat" is "a small vessel for travel on water", to infer that "water" should be more lexically close to "boat". In addition, the hierarchical semantic network from WordNet also provides information like "car" and "boat" are tools of transportation.

Image depository resource. The recent trend of online social networking resulted in many sharing sites. One such online collective image resource is Flickr [Flickr]. The contributors of this website often upload pictures for sharing with meaningful tag descriptors. These tags which describe the images are initially meant for indexing and searching. However, recent research highlighted that such tagging knowledge can also provide useful co-appearance information [Neo07]. Intuitively, by making use of the mutual information between tags, it is possible to guess how likely visual objects can coexist.

Figure 1.1 Retrieval results from Flickr

For example, statistics from Flickr's tags show that "blue, cloud, sunset, water" are the four

most frequently occurring tags with "sky" as in Figure 1.1. It is therefore reasonable to assume that these four visual concepts are more likely to coexist with "sky" than other concepts. This important knowledge can help in improving inference and retrieval.

Parallel news resource. Text articles and news wires are some of the most widely utilized external resources by the research community to supplement retrieval. As news video has an occurrence date, it is reasonable to assume that locating parallel news from external news archives can be carried out without much deterrence. The two most widely used methods to gather news articles are: (a) through online news search engine such as Google [Goog] and (b) newspaper archives. One of the uses for these news articles is for query expansion. This is done by inducing words which have high mutual information with the original query terms. In addition, information from news articles does not have the transcription errors which often appear in comparison to the speech transcripts or closed captions. We can thus leverage this important information to predict missing entities in the speech transcript through an event-based approach.

News blog resource. The next resource which we employ is information from news blogs. This new media has recently attracted tremendous attention from various communities. The rise of blogs is fueled by the growing mass of people who want to express their views and ideas on events. The events they commented on range from their everyday life, current news, animal rights issues, to rumors on celebrities. When a particular high impact event happens, there is usually a sharp rise in "*web activity*" (measured by the number of posted articles) on that event and its related topics. One example is the "*capture of Saddam Hussein*", which triggered a huge number of blog postings and news articles relating to him in December 2003. According to this phenomenon, implicit correlation of the

occurrence and its importance can be derived from the topic's "*web activities*".

1. 2 News Video Retrieval and Question Answering

Retrieval or "search" is the process of finding sets of documents which have high relevance with respect to given queries. This is usually done by estimating the document's relevance against the set of features representing the documents and the query. In traditional text retrieval, the document relevance may simply mean the amount of overlap between keywords and their relationship in the query and in the documents. As we advance from the retrieval of textual data to multimedia data, we observe that queries may not only be consisting of text, but are accompanied by other modalities such as image, audio or video samples. Some examples of available commercial retrieval systems are Google and MSN, which allow users to search for documents, images and even video based on a text query. Other, research oriented, retrieval systems from IBM [Amir05], Informedia [Haup96], and MediaMill [Snoe04] further allow users to supply a text query with multimedia samples during retrieval.

From text-based search using the speech transcripts in the early days, news video retrieval had incorporated the use of low-level video features [Smit02] generated from different modalities, such as the audio signatures from audio stream, or the color histogram and texture from the visual stream. Most existing systems rely solely on the speech transcripts or the closed captions from the news video sources to provide the essential semantics for retrieval as they are reliable and largely indicative of the topic of videos. However, textual information can only provide one facet of news content and offer semantics pertaining to its story context. There are many relevant video clips that might not

carry the relevant text in the transcript and will not be retrievable. In addition, the outputs from an automated speech recognizer and optical character recognizer are not perfect and often contain many wrongly recognized words.

To further improve the accuracy and granularity of video retrieval, some recent research efforts focus on developing specialized detectors to detect and index certain semantic concepts or high level features. High level features denote a set of predefined semantic concepts such as: (a) visual objects like cars, buildings; (b) audio-concept like cheering, silence, music; (c) shot-genre in news like political, weather, financial; (d) person-related features like face, people walking, people marching and (e) scenes like desert, vegetation, sky. The task of automatic detection of high level features has been investigated extensively in video retrieval and evaluation conferences such as TRECVID [Trecvid]. In recent years, researchers [Wu04, Yang04, Yan05] advanced in the development of such detectors, and a large number of high level features can be inferred from the low-level multi-modal features with a reasonable detection accuracy.

While the aim of retrieval is to discover highly relevant documents, question answering can be regarded as a form of precise retrieval which attempt to understand the user's query to locate *exact* answers in which the user is interested. One such example is *"Who was the President of the United States in 2005?"* which requires the exact answer *"George Bush"*. However, an exact precise answer is not useful in video as it is inappropriate to give a short meaningless utterance. For example: it is better to return the whole segment *"Beijing is chosen to be the city hosting Olympic 2008"* rather than just "Beijing" for the query *"Which city will host the 2008 Olympics?"* In short, video question answering requires a good summary. Hence, the problem is different from text-based

question answering. It is also observed in [Lin03] that users show a preference for reasonable semantic units rather than simpleton answers. We conjecture that it would be more applicable for news video since the user can see the enactment in the form of footage while obtaining the information they need.

A user query can generally come from a broad range of domains. In particular, this thesis deals with semantic queries on news video, which aim to find high-level semantic content such as specific people, objects, and events. This is significantly different from queries attempting to find non-semantic content, i.e. *"Find a frame in which the average color distribution is grey"*. [Smeu00] categorized generic searchers into three categories. The first category of users has *no specific interest* but would like to gather more information about latest trends or interesting happenings. The second type of users *knows what they want* and perform an arbitrary search to retrieve documents satisfying their information need. The third kind of users are the *information experts* which require complete information on what they need.

The objective of this work is to provide effective retrieval and question answering to support these users by leveraging computation power to reduce the huge manual annotation efforts. Most of the experiments in this work are carried out based on using *heterogeneous multimedia archives* [West04], which allow huge variability on the topics of multimedia collections. Two examples for heterogeneous multimedia archives are news video archives and video collections downloaded from the Web. This contrasts *homogeneous multimedia archives* collected from a narrow domain, e.g., medical image collections, soccer video, recorded video lectures, and frontal face databases.

1. 3 Proposed Event-based Retrieval Model

For the features from news video as well as the external resources, it is essential to develop principled combination approaches to support precise retrieval. In this thesis, we present our event-based news video retrieval model as shown in Figure 1.2. The framework: (a) represents features at story level from news video to model news events; (b) combines online parallel news and the news video stories for event-based clustering; (c) utilizes the discovered hierarchical structure with other multimodal resources and collective statistics as facets of information relating to an event; and (d) provides advanced query analysis and retrieval to support key event discovery for topic retrieval and video question answering.

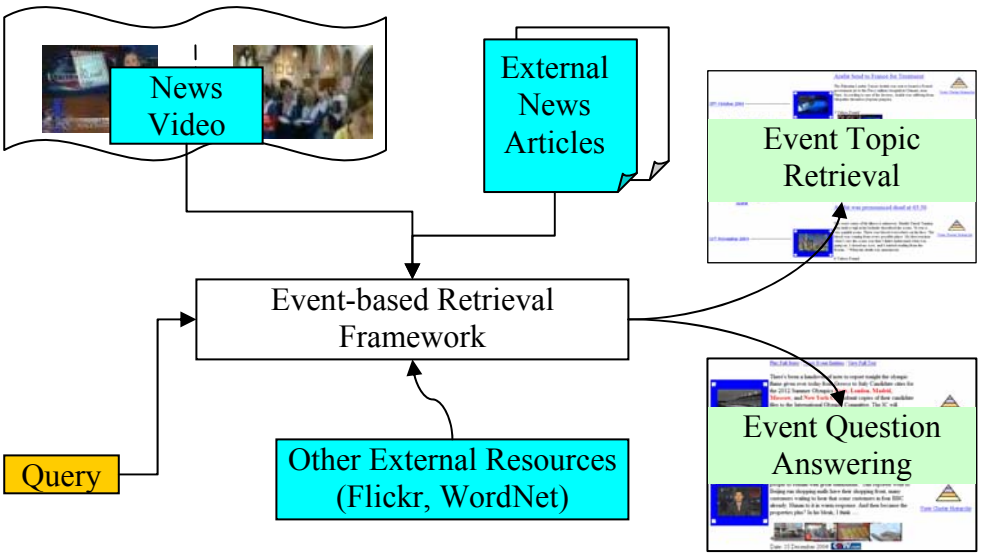


Figure 1.2 Overall Event-based Retrieval Framework

1. 4 Contributions of this Thesis

The contributions of this thesis can be summarized as follows. First, this thesis discovers and describes how external knowledge can be used in supporting various parts of

the event-based retrieval model. In particular, the four proposed resources are language resource, image repository resource, parallel news resource and news blog resources. Several novel approaches are proposed in this thesis, e.g. temporal hierarchically clustering of multi-source news articles and video information based on event entities; blog analysis for key event detection; and combining the language resource and image repository for inference of query high-level features in a query dependent manner.

Second, this thesis presents a news video retrieval framework which combines diverse knowledge sources using our proposed event-based model. This event model integrates multiple sources of information from the original video as well as various external resources. The proposed event-based model has been shown to be robust and effective in retrieval and question answering in the search task of the TRECVID conference. The approaches are evaluated with multiple large-scale news video collections, which demonstrate promising performance.

The thesis is organized as follows. Chapter 2 provides the literature review of related works in the field of text retrieval and multimedia retrieval. It also provides background of work done in text question answering and the use of external knowledge for retrieval. Chapter 3 presents the system overview highlight in contributions in this thesis. Chapter 4 provides the essential background work for video processing. Chapter 5 discusses how multimedia news video is modeled for event-based retrieval. Chapter 6 describes the used query analysis and retrieval process in particular to the proposed event model. Chapter 7 shows the experimental results on large-scale video news collections. Finally, Chapter 8 concludes the thesis and envisions the future of multimedia information retrieval.

Chapter 2

Literature Review

Information retrieval (IR) is the science of searching for specific and generic information in documents, metadata that describe documents, and databases, including the relational stand-alone databases or hyper-text networked databases such as the World Wide Web. An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines. Most IR systems compute a numeric score on how well documents in the database match the query, and rank the documents according to this value. Many universities and public libraries use IR systems to provide access to books, journals, and other documents. Web search engines such as Google, Yahoo search or Live Search (formerly MSN Search) are the most publicly visible IR applications.

The ability to combine multiple forms of knowledge to support retrieval has shown to be a useful and powerful paradigm in several computer science applications including multimedia retrieval [Yan04, West03], text information retrieval [Yang03b], web search [Cui05, Ye05], combining experts [Cohe98], classification [Amir04] and databases [Tung06]. In this Section, we first review some related approaches in the context of text retrieval and multimedia retrieval, followed by reviewing related work from other research areas such as the use of external knowledge and event based retrieval.

2. 1 Text-based Retrieval and Question Answering

Text retrieval is defined as the matching of some stated user query against a set of free-text records. These records could be any type of mainly unstructured text, such as newspaper articles, real estate records or paragraphs in a manual. User queries can range from multi-sentence full descriptions of an information need to a few words. Text retrieval is a branch of information retrieval where the information is stored primarily in the form of text. In recent years, people tend to relate text retrieval directly to search engines as they help to minimize the time required to find information and the amount of information that must be consulted, akin to other techniques for managing information overload. Ranking items by relevance (from highest to lowest) reduces the time required to find the desired information. Probabilistic search engines rank items based on measures of similarity and sometimes popularity or authority. Boolean search engines typically only return items that match exactly without specific order.

One of the most prominent evaluation benchmarks on text processing is the Text REtrieval Conference (TREC) [Trec]. This conference supports research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. In particular, one of the tracks in TREC, the Question Answering track aims to foster research on systems that retrieve answers rather than documents in response to a question. The focus is on systems that can function in unrestricted domains. The target of search will include people, organizations, events and other entities in three types of questions namely: factoid, list and definition questions. Factoid questions, such as "*When was Aaron Copland born?*", require exact phrases or text fragments as answers. List questions, like "*List all works by Aaron Copland*", ask for a list

of answers belonging to the same group. The third type of questions is the definition questions which expect a summary of all important facets related to a given target. For instance, "*Who is Aaron Copland?*" To answer such a question, the system has to identify definitions about the target from the corpus and summarize them to form an answer.

The state-of-the-art question answering systems have complex architectures. They draw on statistical passage retrieval [Tell03], question typing [Hovy01] and semantic parsing [Echi03, Xu03]. In statistical ranking of relevant passages, to supplement the sparseness in a corpus, current systems also exploit knowledge from external resources, such as WordNet [Hara00] and the Web [Bril01]. Given the statistical techniques employed, the techniques focus on matching lexical and named entities with question terms. As such, it is often difficult for existing question answering systems to find answers as they share few words with the question. To circumvent this problem, recent work attempts to map answer sentences to questions in other spaces, such as lexico-syntactic patterns. For instance, IBM [Chu04] maps questions and answer sentences into parse trees and surface patterns [Ravi02]. [Echi03] adopted a noisy-channel approach from machine translation to align questions and answer sentences based on a trained model.

Question answering research has been on-going for more than two decades and its accuracy stands at 70% as published in TREC. To handle news video question answering appropriately, it is important to leverage the know-how from prior works especially in text based question answering as speech transcripts are essentially text. However, the processing of speech transcripts might need different measures as they are usually imperfect. It is therefore necessary to make suitable modifications and adaptations must be applied so as to combine the other available modal features from news video.

2. 2 Multimedia Retrieval and Query Classification

Unlike text retrieval, challenges faced by retrieving multimedia data are much more complex as we face limitations in finding semantic features. It is therefore necessary to apply appropriate techniques in query analysis and fusion strategies so as to handle retrieval of such data. In addition, it is also important to derive usable semantics from the low level non-semantic features. Various studies such as [West03] have shown that retrieval models and modalities can affect the performance of video retrieval. [West03] adopted a generative model inspired by a language modeling approach and a probabilistic approach for image retrieval to rank the video shots. Final results are obtained by sorting the joint probabilities of both modalities. In general, two distinct retrieval strategies can be seen in the multimedia community: one that uses generic retrieval (query class independent) while the other fuses features accordingly to query properties (query class dependent).

In query class independent retrieval, the system employs the user's queries to find relevant shots or segments using the same generic search algorithm or fusion parameters. The video retrieval system proposed by [Amir03] applied a query class independent linear combination model to merge the text/image retrieval systems, where the per-modality weights are chosen to maximize the mean average precision score on the development data. Other retrieval systems such as [Gaug03] ranked the video clips based on the summation of feature scores and automatic speech retrieval scores, where the influence of speech retrieval is four times that of any other feature. [Raut04] used a Borda-count variant to combine the results from text search and visual search. The combination weights are pre-defined by users when the query is submitted. However, until recently most of the multimedia retrieval systems use query class independent approaches to combine multiple knowledge sources.

This has greatly limited their flexibility and performance in the retrieval process [Yan03]. Instead, it is more desirable to design a better combination method that can take query information into account without asking for explicit user inputs.

Recently, query class dependent combination approaches [Yan04, Chua04] have been proposed as a viable alternative to query class independent combination, which begins with classifying the queries into predefined query classes and then applies the corresponding combination weights for knowledge source combination. In [Yan04], they followed a conventional probabilistic retrieval model and framed the retrieval task using a mixture-of-expert architecture, where each expert is responsible for computing the similarity scores on some modality and the outputs of multiple retrieval experts are combined with their associated weights. Four classes are defined: Object, Scene, Person and General. The text features provide the primary evidence for locating relevant video content, while other features offer complementary clues to further refine the results. However, given the large number of candidate retrieval experts available, the key problem is the selection of the most effective experts and learning the optimal combination weights. The solution is an automatic video retrieval approach which uses query-class dependent weights to combine multi-modality retrieval results.

In this work, we make use of query class dependent retrieval [Chua04, Neo05] as the basis for fusion of multimodal features. Crucially different from [Yan04], our query-classes follow the genres of news video (e.g. sports, politics, finance, etc). We are among the first few groups to leverage the idea of query classification. Experimental evaluations have demonstrated the effectiveness of this idea, which have then been applied in the best-performing systems of TRECVID search task from 2004 to 2006. This is further validated

by many follow-on studies [Chua05, Hsu05, Kenn05, Huur05, Yuan05] which shows positive usage of query classification. For example, [Huur05] suggested it is helpful to categorize the queries into general/special queries and simple/complex queries for combination. [Yuan05] classified the query space into person and non-person queries in their multimedia retrieval system.

2. 3 Multimodal Fusion and External Resources

An alternative approach for multimedia retrieval is to use text-based structural data retrieval techniques to search structural data representation that includes all the information of textual features and semantic concepts. The "Multimedia Content Description Interface" (MPEG-7) [Smit03] is the most widely adopted storage format for video retrieval. A number of successful video retrieval systems have been built upon the MPEG-7 representation. For instance, a MPEG7 framework to manage the data in audio-visual representation is proposed in [Tsin03]. The annotation is based on fixed domain ontology from TV-Anytime and the retrieval is restricted to querying the metadata for video segments. [Grav02] proposed an inference network approach for video retrieval. The document network is constructed using video metadata encoded using MPEG7 and captures information about different aspects of video. To provide more semantic and reasoning support for the MPEG7 formalism, a framework for querying multimedia data using a tree-embedding approximation algorithm has been proposed [Hamm04]. Generally speaking, the knowledge sources provided by textual features, image features and semantic concepts are treated differently in these text-based approaches. However, this style of retrieval requires most features to be meta-indexed first which may not be suitable as this will require huge annotation efforts.

The paradigm of lessening human annotation efforts triggers the use of extracting high level semantic concepts from multimedia streams automatically. The community has investigated this in part by developing specialized detectors that detect and index certain high level features (e.g. cars, faces or buildings). With this methodology, search can be carried out by combining multiple detection models; or combining the detection models with different underlying features; or combining the models with the same underlying features but different parameter configurations. Among them, the simplest methods are those fixed combination approaches. The IBM group [Amir06] fused a series of low-level features and high level features based on two learning techniques. Their system maps query text to high level features models by co-occurrence statistics between speech utterances and detected concepts as well as by their correlations. The MediaMill group [Snoe06] further extended the LSCOM-lite set by adding more HLF (total of 101) to support the same task. Other top performing interactive retrieval systems from Informedia [Yang06] and DCU [Fole05] have integrated the use of high level features in their retrieval. Even though the detection rates of high level features are relatively low, recent results show that they can be used to supplement text in improving multimedia retrieval performance [Trecvid].

However, most of the prior works does not leverage semantic inference of the text query to available high level features during retrieval. The inference step is important as the set of high level features is limited. To cater to a wider range of queries, we propose to use external knowledge such as WordNet to relate text queries to high level features through the use of its glosses. The use of WordNet has been widely discussed but primarily on the use of its semantic network. In our work, we focus on using the gloss as they can sometimes provide more relevant details in terms of descriptions than the hierarchical structure.

2. 4 Event-based Retrieval

News stories are depictions of real-life happenings. In simpler terms, news video stories can be seen as materials consisting of both text and visual information of a real life event. Intuitively, the text-terms like the persons' names, locations and activities is made up from the actual event entities in a real life happening. Visual revelations from the visual stream of news video constitute the event scene. This morphology of news video retrieval is similar to what is known as an event-based retrieval in text retrieval.

In fact, event-based structured retrieval has been shown to be effective in retrieval and question answering [Yang03b]. They also observed that a question answering event shows great cohesive affinity to all its elements and the elements are likely to be closely coupled by this event. Normally, the question itself provides some known elements and asks for the unknown element(s). Thus, it is possible to make use of these known elements to induce unknown elements in a closed temporal domain. To tackle the problems of insufficient known elements and inexact known elements, they [Yang03b] modeled the Web and linguistic knowledge to perform effective question answering. The grouping of events by their relative similarity and differences also helps in tracking events across time. This is similar to topic detection and tracking (TDT) [Alla98]. TDT attempts to use the lexical similarity of the document text to generate coherent clusters, in which elements in the same cluster belongs to the same topic. If such topic/event structures are available, it can provide excellent partial semantics for retrieval as well as news video threading [Hsu05]. However, TDT on text faces natural language processing issues like word sense ambiguity and it is even more challenging for news video since the speech transcripts are erroneous.

2. 5 Summary

With the advent of the Internet, the Web has grown to an enormous knowledge repository and archives more information than any library on the planet. Many systems therefore utilize the vast Web resources to enhance retrieval and question answering. This is especially so in text retrieval systems such as [Bril01] that uses the collective search statistics from the Web for calculating mutual information of terms.

In recent year, there is a massive growth of social networks and online folksonomies. These sources of new media provide valuable knowledge which can be leveraged during retrieval. For example, lexical similarity from WordNet dictionary does not necessarily means visual similarity, and it is thus necessary to provide other resources to better measure visual similarity. In particular, we propose the use of information from photo sharing sites like Flickr [Flic]. Flickr allows users to upload images with tags and these tags can be useful in providing information on the co-appearance quality between visual objects.

To handle limitations in TDT, we propose to perform clustering using event entities. High level features are also used in the clustering step as they can be indicative of the topics or events. For example, the presence of high level features like water-bodies and buildings in the video scene may indicate a flood event. We supplement the clustering space with external news articles from the Web to provide a semantic bridge during clustering.

Besides using external parallel news for clustering, we also propose to obtain event "interestingness" from the Web by considering "web activities" from news blogs. Mapping the video news stories into events allows us to measure how much of web activities are centering on a particular topic/ event and thus providing an estimate of its interestingness to general users.

Chapter 3

System Overview and Research Contributions

This Chapter provides an insight to the event-based retrieval model with details on research contributions for the thesis. The proposed system consists of two main parts: (a) offline feature extraction and event representation and (b) real-time topic retrieval and question answering. The overall architecture is shown in Figure 3.1.

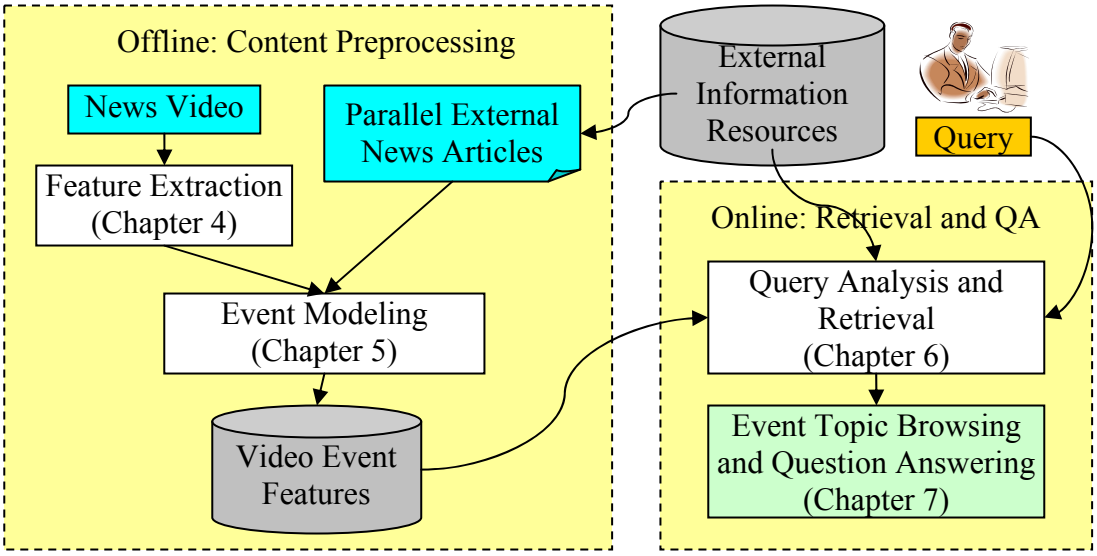


Figure 3.1 System Overview

3.1 Content Preprocessing

In the preprocessing stage, the system takes in raw news video files in digital format with or without meta-data files. The first step would be the primary feature extraction that involves the extraction of a variety of low-level video features from different modalities,

e.g., audio, speech and visual frames. Many prior works [Haup96, Wact00, Amir03] centered on utilizing useful low-level visual features, text features, audio features, motion features and other metadata at shot level. One of the main reasons why video is often represented at shot level is that the shot is the smallest semantic unit next to a video frame and state-of-the-art shot boundary detectors are excellent. In recent years, news video retrieval has incorporated the use of high-level features for specific objects or phenomenal (e.g., cars, fire, and applause), often organized in the form of a large hierarchical concept ontology. A well-known example is the Large-scale Concept Ontology for Multimedia (LSCOM) [Lscom], which contains approximately 1000 concepts that can be used for annotating videos. This thesis provides combinatorial approaches in combining outputs from multiple SVM detectors to enhance detection accuracy for improving the retrieval quality. Chapter 4 of the thesis will provide the basic background and existing techniques for content-based video feature extraction that is required for the event-based retrieval model.

Following the video feature extraction process, Chapter 5 will cover the **first major contribution** of the thesis, the event-based modeling of video features. The event-based retrieval model extended video features on a story level basis by using the discovered story boundaries. This approach is rational and intuitive since story boundaries are determined based on a change in news topic. The main intuition for linking to events is to leverage innate associations among elements of the events. It is then possible to leverage mutual information [Kenn89] to group known event elements together, and even subsequently predict missing entities during retrieval. The thesis moves on to describe the proposed temporal hierarchical *k*-means clustering on the video corpus to obtain homogenous grouping of news instances. However as video features are often noisy with necessary

entities related to an event missing, a systematic approach based on using external resources in a temporal fashion is proposed. The thesis introduces two adaptations to traditional data clustering by: (a) multi-stage clustering which uses different set of features for clustering at different hierarchy level and (b) imposing a temporal partitioning and subsequent recombination in the clustering space. The adapted temporal hierarchical *k*-means clustering resulted in better quality clusters and can be carried out efficiently in a large video corpus.

In addition to getting video features from within the original source, the thesis further provides intuitive methods to source for other multimodal features from unrefined collective information online. In particular, it proposes a novel approach to obtain event "interestingness" through news blog sites. This event interestingness can then be leveraged during retrieval to support topic evolution browsing.

3.2 Real Time Query Analysis, Event Retrieval and Question Answering

Besides modeling features in the point of view of event space, interpreting the user's query correctly is also crucial for an effective retrieval. In particular, Chapter 6 focuses on the **second major contribution** of the thesis that is on leveraging external resources in event query analysis, retrieval and browsing. The query analysis module extracts a series of query features like query-terms, query-high-level-features and query-class. To gather additional query-terms, we incorporate the use of parallel news in a temporal fashion during query expansion to infer additional terms. The rationale for using a set of temporal close news articles is to preserve context and reduce noise during query expansion. The query-High-level-feature is useful in relating the importance of available high level feature to the query so as to leverage those relevant high level features during retrieval. To measure this

importance appropriately, we intuitively combine various external resources such as Wordnet and Flickr for lexical and visual co-occurrence similarities respectively. Our strategy improves existing work which uses WordNet by further considering word glosses as they sometimes provide visual descriptions that are not available in the WordNet lexicon hierarchy. Flickr on the other hand is used to calculate the co-appearance quality using mutual information from the image tags contributed by users.

During retrieval, the system facilitates both story level retrieval and shot level retrieval according to the user's query. For shot retrieval, query-classification dependent fusion is used to combine various modal features at the shot level. Query-class is necessary as different queries have different characteristics and therefore require different features as evidence. For example, a person-directed query will likely rely more on the video captions as compared to a sports query. To further improve precision, a round of pseudo relevance feedback is done using the top retrieved shots.

The thesis then moves on to describe two applications which leverage on the event-based retrieval framework. The first application is event topic browsing. At present, news video search engines display lists of candidate results arranged in order of relevance to these users (see for example, commercial sites such as www.streamsage.com). Such an arrangement might be good for collecting data related to the topic, but may be too data-overwhelming for most users. Leveraging on the clustering results from the event-based model, we generate a *query topic graph* that is a graphical structure containing relevant materials to a user's query. This query topic graph makes use of the event cluster information to present to the users a more structured output during browsing. This can be useful in locating related video events or getting a topic overview. For example: given a

query on "*Arafat*" on November 2004 where Arafat has just passed away, it would be good to present an overview of reports arranged in a chronological order with sub-topics determined by level of interestingness like "Arafat is hospitalized", "Arafat has fallen into coma", and "Arafat is pronounced dead by the Palestinian officials". This kind of grouping of search results is similar to text clustering done in a commercial search system named Vivisimo [Vivi]. The added advantage of such a presentation is that it is capable of showing the key stages of the news topic based on interestingness when arranged chronologically.

The second application is event question answering. In contrast to event topic retrieval, event question answering aims to find exact multimedia answers targeting at *specific aspect* of an event. A user initially looking for news on "*Arafat*" may have follow-up questions like "*when was Arafat hospitalized?*", "*which hospital did he go to?*" Such information needs require a finer interpretation of the user's intention. We employ query typing [Yang03] that is widely practiced in text question answering to predict the plausible answer type. A word density-based ranking is then used to select the most possible answer candidate from the news video stories retrieved from the query topic graph. Beside the text-oriented questions above, it is also possible that users are interested in specific visual details like "*shots containing Arafat*" or "*shots on Arafat's funeral*". This type of questions will largely depend on the visual stream for visual evidence. To enhance the recall of locating visual answers, we further make use of the event-based model by expanding the query topic graph to find other relevant news video events.

To understand the effects of the proposed event model and techniques, a full set of experiments following the automated search task in TRECVID [Trecvid] is carried out in Chapter 7. The thesis then concludes with possible future work in Chapter 8.

Chapter 4

Background Work: Feature Extraction

The first step is to gather sufficient discriminating features so that the system is able to identify relevant information from the mass of irrelevant ones during retrieval. As video is a continuous stream of multimedia information, it is necessary to determine a suitable unit for its content representation ideally capturing individual events. The most widely employed unit in the multimedia community is a shot, defined as individually recorded segments following the start of camera recording to a pause or stop. State-of-the-art shot boundary detection system is excellent and efficient [Hua02, Quen04, Pete05]. Once the shot boundaries are found, key-frames can then be generated for static visual feature extraction. Representing videos in terms of shots is visually intrinsic but may be inconsistent with other modalities such as the video speech, as story narration may not coincide directly with shot boundaries. A single continuous speech made by the same speaker in the same topic can span multiple shots, causing fragmentations in speech if it is segmented based on shot boundaries which in turn results in meaningless speech segments. To handle such deficiency, techniques based on speaker-change or story change are being developed [Adco02, Chai02, Chri02, Hsu05]. The story offers a more appropriate unit for general news retrieval as it provides better coverage of an event. In this chapter, we will describe in detail the preprocessing of content and features for news video.

4. 1 Shot Boundary Detection and Keyframes

The first step in the content processing of the news video is the detection of shot boundaries. Digital video is organized into frames - usually 25 or 30 per second. The next largest unit of video both syntactically and semantically is called the shot. A half-hour video, in a television program for example, may contain several hundred shots. A shot is produced by a single run of a camera from the time it was turned on until it was turned off. Generally, there are two kinds of shot boundaries: cuts and gradual transitions. Shot boundary detection is an essential step in segmentation of video data. Most methods [Trecvid, Chai02] are based on frame comparison (dissimilarity measure) such as pixel-by-pixel frame comparison. This gives good results but induces a very high complexity and it is not robust to noise and camera motion. There are also methods [Quen04, Amir04] that represent frame content by histograms and vector distance measures. This produces a good frame dissimilarity measure, but histograms lack spatial information, which needs to be compensated with local histograms or edge detection.

[Quen04] used direct image comparison for cut detection. In order to reduce over-segmentation, frames are compared together after motion compensation and a separate camera flash detection is also used. IBM proposed a system [Amir04] that employs a combination of three-dimensional RGB color histograms and local edge gradient histograms. Adaptive thresholds are computed by using recent frames as reference. MSR-Asia system [Hua02] uses global histograms in the RGB color space. These systems are able to produce a detection accuracy of about 90% or more. In this thesis, we utilize the shot boundary detector and key-frame extractor from [Pete05]. Even though we treat shots as the basic unit of content representation, it is an open research issue on how to effectively model

the temporal contents of a shot appropriately based on its features. At present, researchers like to associate a shot with its frame(s) that is/are most informative, known as keyframes. Generally the keyframe extraction process is integrated with the processes of segmentation. Each time a new shot is identified, the key-frame extraction process is invoked, using parameters already computed during segmentation.

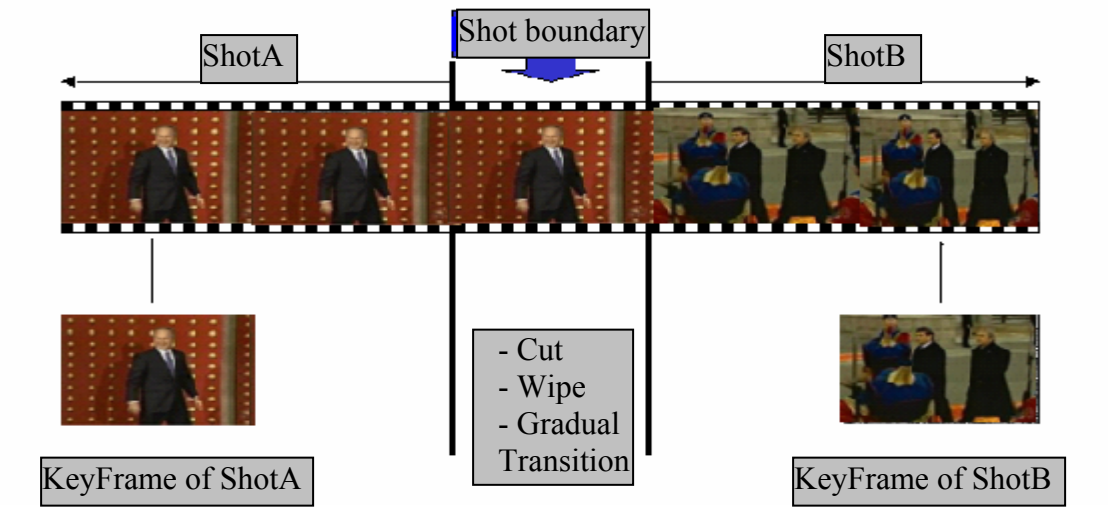


Figure 4.1 Shot detection and keyframe generation

4. 2 Shot-level Visual Features

Given the keyframes, we then extract the following visual features as the representative features of the shot. Note that there could be multiple keyframes in a single shot so as to provide more semantics about the shot.

Color, texture, edge and motion features. Color is the most basic attribute of visual contents. Forms of color representation include dominant color, color histogram and color moments. Edge feature represents the spatial distribution of the image. It consists of local histograms describing the distribution of edges in directional or non-directional manner. The texture descriptor is designed to characterize the properties of texture in an

image (or region), based on the assumption that the texture is homogeneous – i.e. the visual properties of the texture are relatively constant over the region. Motion feature captures the intuitive notion of *"intensity of action"* or *"pace of action"* in a video segment. It is also used in extracting the basic camera operations: fixed, panning, tracking, tilting, booming, zooming, dollying, and rolling. Most image matching techniques employ one or more such features. One problem with using too many visual features is the curse of dimensionality, which will imply bigger latency during retrieval. In order to allow for real-time searching, we restrict the feature size to 116-dimensions for each key-frame as shown in Table 4.1.

Table 4.1 Low level features extracted from key-frame (116 dimensions)

Color Moments	**Edge**	**Motion**
3X3 block with 1st, 2nd and 3rd color moments (27 Dimensions)	Normalized Local edge histogram (80 Dimensions)	1 Global, 8 Directional (9 Dimensions)

Scale-invariant feature transform (SIFT) features. SIFT features are invariant to image scale, rotation, and partially invariant to changing viewpoints as well as change in illumination [Lowe04]. SIFT algorithms transform image data into scale-invariant coordinates relative to local features. Such feature representations are thought to be analogous to those of neurons in the inferior temporal cortex, a region used for object recognition in primate vision.

We employ SIFT in near duplicate key-frame detection that detects the same or duplicate scene but with slightly different visual appearance. The reason for such a visual difference is due to the geometric and photometric changes caused by the variance of video shooting angle, lighting condition, camera sensor or video editing process. Detecting near duplicate key-frames in a video corpus is important as it helps to build up the linkage of

relevant news stories across different TV news channel, language, and time. However, it is known that the computational cost for SIFT especially across large image or video datasets can be high. In this thesis, we employ a fast algorithm [Zhen06] implementation that makes use of a pre-clustering step to group similar key-frames in a video corpus together and then perform near duplicate key-frames within each individual clusters. This clustering step is based on a set of globally invariant image features like the auto-correlogram of the transformation of color intensities. The transformation makes the color features invariant to illumination change by normalizing color intensity with its average intensity and variance. We then apply SIFT-based image matching [Miko05] within the cluster to determine which are the images are near duplicates of one another. This result is then stored in an (|S| × |S|) matrix D_s, providing information of near duplicate keyframes in the corpus.

$$D_s = \begin{Bmatrix} 1 & 0 & 1 & \dots \\ 0 & 1 & 0 & \dots \\ 1 & 0 & 1 & \dots \\ \dots & \dots & \dots & \dots \end{Bmatrix} \tag{4.1}$$

The matrix D_s indicates that shot_1 and shot_3 contain near duplicates keyframes (v=1).

Video captions. This feature is obtained through the use of a video optical character recognizer (VOCR) on video frames. Video captions can provide two sources of information. The first source of information is the parallel speech by the narrator. This source of information can be leveraged to supplement the speech transcripts. The second source of information is the identity caption. It is common to see names of news subjects appearing in the caption during news reporting. In addition, natural disaster events such as floods, tsunamis, etc, usually have the corresponding wordings of the event and location displayed together with the live-reporting shots or footages. We employ the CMU VOCR

[Chen04] in this work to detect captions. The system predicts possible positions on the frame that may contain text and then make use of a character recognizer to detect the wordings.

4. 3 Speech Output

One of the key features that enable the efficient retrieval of news video is the speech transcripts. The use of speech transcripts can suggest the approximate story context and help to bridge the gap from low-level audio features to the semantic intepretation. Based on past experiences, a good commercial automated speech recognizer (ASR) [Gau02, Van02] can achieve a detection accuracy of about 70-75%. For multilingual non-English news video, there is an additional step of translating the non-English transcripts to English.

In addition to ASR text information, speaker change information is also normally available through the automated speech recognizer. Such information is useful in generating pseudo sentences and providing a smaller coherent retrieval unit. In particular, machine translation is also based on phrase units arising from either a change in speaker, pauses or silence. Previous experiments in news video retrieval [Adco05, Chua05] showed that speaker-change text units provide better context than the shot-based text units in retrieval.

4. 4 High Level Feature

The image/video analysis community has struggled to bridge the semantic gap from low-level feature analysis (color histograms, texture, edge) to semantic content description of video. For query such as *"Find scenes containing a car"*, it is currently not possible to automatically determine which low level features in the image or video are discriminative to

identify the main object cars with high accuracy. To overcome the semantic gap, one approach is to utilize a set of intermediate semantic concepts known as high level features (HLFs) that can be used to describe frequent visual and audio content entities in video collections. The set of HLFs defined in the recent TRECVID conference include car, person, explosion, fire, etc [Trecvid]. A HLF detection algorithm is made up of three major building blocks: low-level feature extraction, uni-modal feature-based learning and multi-modality fusion. Based on the aforementioned low-level features, the detector can be constructed using statistical learning algorithms such as the support vector machines (SVMs) [Joac98, Amir04, Chua04]. Built on the structural risk minimization principle, SVMs aim to seek a decision surface that can separate the data points into two classes with a maximal margin between them.

IBM [Amir04] studied the min, max and un-weighted linear combination function for multi-modality and multi-model fusion for HLF detection. [Yang04] specifically considered the problem of detecting news subjects in news video archives by linearly combining the multi-modal information in videos, including transcripts, video structure and visual features. [Snoe5] compared the early fusion and late fusion methods by using SVMs as the base classifiers and meta-level classifiers for fusing text and images. The current performance of detectors can vary greatly, depending on the difficulty of the concept. In TRECVID, the state-of-the-art detectors can fetch a mean average precision (MAP) of about 0.3. Besides relying on a single detector output to support retrieval, it is also possible to combine results from various detectors for the same concept to enhance accuracy. The rationale is as follows: (a) different systems have their own unique way of detecting HLFs, and therefore combining these results or rank lists can have a complementary effect; and (b)

with the growing importance of HLF detection, we can expect more and more such individual detection results or detectors from different groups to be available online. In TRECVID, participants in the search task can make use of the HLF detection outputs from different participating groups in the HLF detection task.

In this thesis, we have identified 50 HLFs that include the 39 LSCOM-lite concepts, and 11 genre and audio concepts determined by us as shown in Table 4.2. The LSCOM-lite is a subset of concepts from LSCOM that are used for evaluation in TRECVID 2005 and 2006. Instead of building our detectors for all the 50 HLFs, we leverage the detection outputs from various systems [Amir05, Fole05, Haup05, Snoe05] from the TRECVID HLF detection task. The detection output of each of the various groups is a ranked list (omitting confidence values) containing a maximum of 2000 shots for each of the 39 LSCOM-lite HLF, with the first shot having the highest confidence. We fuse the multiple rank lists using the following two approaches: (a) score fusion and (b) rank fusion.

The score fusion approach takes into consideration the position of the shot across multiple rank lists as shown in Eqn 4.2. Here, we estimate the HLF detection confidence of a shot by its position in the rank list, although ideally, using the individual system's detection confidence will be more appropriate but this information is not available in the rank list.

$$Conf_{shot}(s \mid HLF_k) = \alpha \cdot Contains_t(s) + (1-\alpha)\sum_t \frac{\max_t Pos_t - Pos_t(s)}{\max_t Pos_t} \qquad (4.2)$$

where *Contains(s)* gives the number of rank lists *t* that have shot *s*, and the second term produces a normalized score in the range of [0..1] that linearly weights the position (Pos_t) for shot *s* on the ranked lists *t*.

Table 4.2 Description of High Level Features (* denotes not in LSCOM-lite)

Type	Description	High Level Feature
Program Genre	Program genre information of news video clip which is useful in describing the properties of news video. This is extracted by employing similar text classifier from [Chua04] over the speech transcripts at pseudo story level.	*Political*, Scientific*, Entertainment, Sports, Weather, Financial*, Disaster, Military* Commercial*, General**
Person-related	These are the high level features suggesting the presence of a face, a known person, type of people or people in action. Face detection is first carried out on the keyframes of the shots. Face recognition or person-X [Zhao06] detection is then subsequently carried out for a set of known news subjects. The algorithm uses a fusion of face information, speech and video captions with external information.	*Face, Person, Corporate leaders, Government leaders, Prisoners, Military, Police, People walking and People marching, Anchor-person*
Audio Genre	Audio information such as audio genre is another important feature for news video. The detection is based on the Mel-frequency Cepstrum Coefficient (MFCC) together with zero crossing rates, centroid and roll-off point energy as features [Jian00].	*Cheering*, Music*, Speech*, Noise**
Scene	Scene high level features describe the relevant news scenes in which the news video is set or filmed. They are generally useful during pseudo relevance feedback as answer shots have intuitive similar background.	*Indoor: Court, office, meeting and studio Outdoor: Buildings, Road, Sky, Snow, Urban, Desert, Vegetation, Mountain and Waterscape, Crowd*
Objects/ Others	These are the objects, abstract concepts or other high level features which are not categorized into the previous four types.	*Flag-US, Airplane, Car, Bus, Truck, Boat, Ship, Maps, Fire,* E*xplosion*

Rank fusion differs from scoring fusion in that it concentrates on the rank rather than the confidence. As confidence from different systems may have different meaning, they may be difficult to be made comparable. Rank list fusion can avoid this problem by only utilizing the relative ranks. In practice, we employ Rankboost [Freu97], a powerful algorithm for combining rank lists or preferences. It has been successfully used in information retrieval, natural language processing and shape localization. RankBoost is similar to AdaBoost [Poli06] or adaptive boosting. They differ in that AdaBoost combines classifiers while

RankBoost combines ranking functions. AdaBoost is adaptive in the sense that subsequent classifiers built are tweaked in favor of those instances misclassified by previous classifiers while RankBoost focus more on returning an optimal rank list. The algorithm for RankBoost is shown in Figure 4.2.

Algorithm RankBoost

1. Given initial distribution P over $S \times S$
2. Initialize $P_1 = P$
3. For $t = 1, \ldots, T$:

- Train weak learner using distribution P_t
  - Get weak ranking $h_t : S \rightarrow \{0, q_{def}, 1\}$
  - Choose $\alpha_t \in \mathbf{R}$
  - Update: $P_{t+1}(s_0, s_1) = \frac{P_t(s_0, s_1)\exp(\alpha_t(h_t(s_0) - (h_t(s_1)))}{Z_t}$

where Z_t is a normalization factor

4. Output final ranking: $H(s) = \sum_{t=1}^{T} \alpha_t h_t(s)$

Figure 4.2 RankBoost Algorithm from [Freu97]

Let $s \in S$ be the set of all shots in the news video corpus. We use RankBoost to build a strong ranking function H to combine the rank lists from T weak ranking functions into a single final rank list based on some *feedback* information. In this case, we treat the T weak ranking functions as rank lists f_n from various detectors. A weak ranking function h_t is given as in Eqn 4.3.

$$h_t = \begin{cases} 1 & if\ f_n(s) \geq \rho \\ 0 & if\ f_n(s) < \rho \\ q_{def} & if\ f_n(s) = \perp \end{cases} \qquad (4.3)$$

where $\rho \in \mathbf{R}$, $q_{def} \in \{0,1\}$ *and* $f_n(x) = \perp$ means that x is unranked by rank list f_n since it is highly possible that a particular shot is not ranked by every system. In this case, a default score will be assigned. The feedback information takes the form of $\phi : S \times S \rightarrow R$, where

$\phi(s_0, s_1) > 0$ means that s_1 should be ranked above s_0 while $\phi(s_0, s_1) < 0$ means the opposite; and zero indicates no preference. Finally, the final ranking function *H* combines all the weak ranking functions h_t to form a final rank list. The use of RankBoost, however, requires some form of feedback to adjust the weights given by various weak ranking functions h_t. It is therefore necessary to supply an initial set of shots that have a high confidence of accuracy for the purpose of feedback. For this, we make use of the top *n (where n=50)* shots obtained from Eqn 4.2. To illustrate the improvement in performance, we compared the MAP (mean average precision) performance of rank lists from various top detection systems, as well as our final rank list employing RankBoost in Table 4.3.

Table 4.3 MAP performance: Comparing the top 3 performing systems (S1, S2, S3, T1, T2, T3) reported in TRECVID 2005 and 2006 with score fusion and RankBoosting (* TRECVID 2006 uses inferred MAP for assessment)

TRECVID2005	**S1 (best reported)**	**S2**	**S3**	**SF**: Score fusion using Eqn. 4.2	**RF**: Rank fusion
MAP	0.351	0.334	0.308	0.398 (+13%)	0.443(+26%)

TRECVID2006	**T1 (best reported)**	**T2**	**T3**	**SF**: Score fusion using Eqn. 4.2	**RF**: Rank fusion
MAP*	0.261	0.222	0.186	0.308(+18%)	0.333(+28%)

From Table 4.2, we observe that the SF and RF runs generally perform significantly better than the single detector runs (S and T runs). This observation validates our earlier hypothesis that combining various detection results can complement one another. In addition, we also see that Rank fusion can achieve about 26% and 28% improvement over the best performing system for 2005 and 2006 TRECVID dataset. The results also show that Rank fusion yields more improvement than score fusion.

4. 5 Story Boundary

The previous few Sections discuss how we represent content at the shot level. In this Section, we will discuss another semantic unit, the story, which can be used to organize digital video and in particular news video. A news story is defined in [Chai02] as a series of related multimedia information centering a particular topic or genre. A news story is defined in [Smea03] as a segment of news broadcast with a coherent news focus which contains at least two independent, declarative clauses. For example, Figure 4.3 shows a series of shot belonging to a news topic depicting "George Bush" and "winter snow".

Figure 4.3 Shots belonging to a single news video story

Many prior works in news video story segmentation employ SVMs to discriminate shots that are likely to contain a story boundary. They mainly use features such as speech from audio channel and other low level visual features. Besides such features, several systems also utilized special characteristics of news video such as *"lead in/lead out"* or *"logo switch"* (a changed in the logo signifies a new story for some news broadcasting stations) or even heuristics like density of news stories. [Hoas04] used SVMs to calculate the distance between each shot vector and the hyper-plane of the SVM to determine the top *N*-shots, where parameter *N* is decided based on the average number of story boundaries in

the development data set, i.e., 20 for ABC news and 36 for CNN news, respectively. [Chai03] developed a two-level story segmentation scheme in which they first performed shot genre classification using low level features to classify a shot into classes like Anchor-person, Live-reporting, Speech Interview, Sports and Text-scene. Subsequently, they utilized the shot genre and other time-dependent features such as the speaker change, scene change and key phrases in a Hidden Markov Model to predict the story boundaries. In addition, they further employed heuristics to decide if "Anchor-person" shots should be further segmented.

In this work, we make use of the story boundaries output provided by [Hsu05b] that uses low level and mid-level features for segmentation. We then enhance their story boundaries output by utilizing detectable "anchor-person" shots from HLF. This is done by fine cutting lengthy segments (>100 seconds) into smaller segment if there are "anchor-person" shots within. The underlying reason is that we prefer over rather than under-segmentation so as to reduce the amount of segments which contains different stories.

Chapter 5

From Features to Events: Modeling and Clustering

Chapter 4 presents the features extracted from news video. In this Chapter, we will describe in detail the modeling of news video at an event level, as well as provide the basis to extend the video features into the event space to support event-based retrieval. Following the definitions of events in text question answering, an *event* is defined as happenings that occur at a given time and place; and a *topic* as a grouping of related events occurring across the temporal domain. In this thesis, a news video event is mapped to a unit of news video story following the story segmentation results. The news video events in the corpus are then grouped using unsupervised clustering. The rational is to obtain a semantically organized news cluster structure that can: (a) provide a cluster-based inference that makes it possible to find relevant news stories even when certain key entities may be missing in the news video feature; (b) provide a basis for inferring additional relevant stories or shots since news video of the same events tend to be similar in terms of context and visual information, and (c) understand the topic flow or evolution by threading events across time. In this Chapter, we will discuss how external news articles are utilized together with the news video stories in a temporal clustering fashion.

5. 1 Event Space Modeling

The relationship between news and events is straightforward. Event is a happening while news is a report on events which are generally important in nature. In particular, video

news usually feature news that is most interesting since airtime is expensive and limited. This is quite different from news articles presented online or in the newspaper, where they provide a less expensive medium for conveying messages to mass media. Typically, we note that news reports contain aspects and entities like: *Location, Time, Subject, Object, Quantity, Action and Description,* etc. The following definition defines a generic news event.

Definition 1. *(news event) An occurrence in real time having describable properties such as time, location, subject, object, quantity, action and description.*

News video inherits the traits of news but also contains additional information in the form of multimedia depictions. This usually includes scenes of the actual event or happening, interviews or speech of the related subjects, etc. This form of information provides information in the multimodal aspects that is not fully conveyable by using only text. In particular we have the following definition for multimedia news.

Definition 2. *(multimedia news event) An occurrence in real time having describable properties such as time, location, subject, object, quantity, action and description with intrinsic and extrinsic visual, audio and related media.*

We consider an event as a distinct point in the multi-dimensional event space. Figure 5.1 presents a simplified 4-D view of an event containing information about the time, location, action and possible HLFs. One example of such a 4-D event derive from news video can simply be a footage of President George Bush giving a speech at the White House on the 3rd of Nov regarding presidential election. (time: 3 Nov; location: White House; action: election, speech; HLF: George Bush, person)

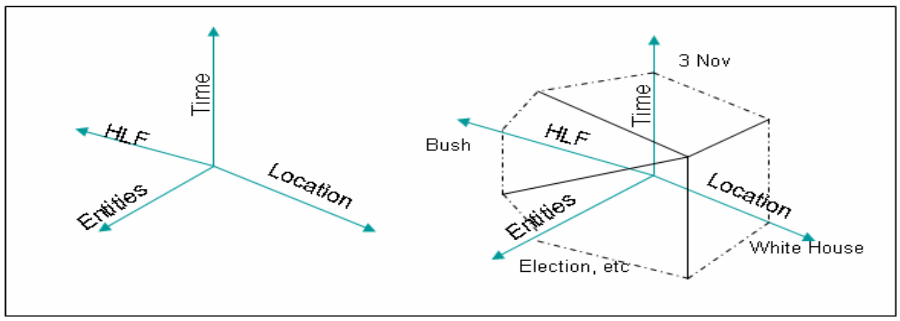


Figure 5.1 Representing a news video in event space

The amount of event information obtainable from news video determines how well we can perform the retrieval. For example, if the query is directed at finding news related to the election of President George Bush, the text information will suffice. However if the query is more specific like "Find shots containing George Bush speaking at The White House", visual information in addition to text will be necessary. It is clear that the primary source of event information comes from speech transcripts. From intuition, we know that it is likely that information such as location, time, subjects and type of activities can be gathered from the speech transcript. However, as transcripts are seldom perfect with many unavoidable translation errors from the machine translating process, it is not sufficient to base it on speech transcripts alone. To gather more information, we combine other multimodal features available from the video source, as well as employ external resources to provide additional facets of information for a news video event. Figure 5.2 shows the overview of event entity extraction from news video.

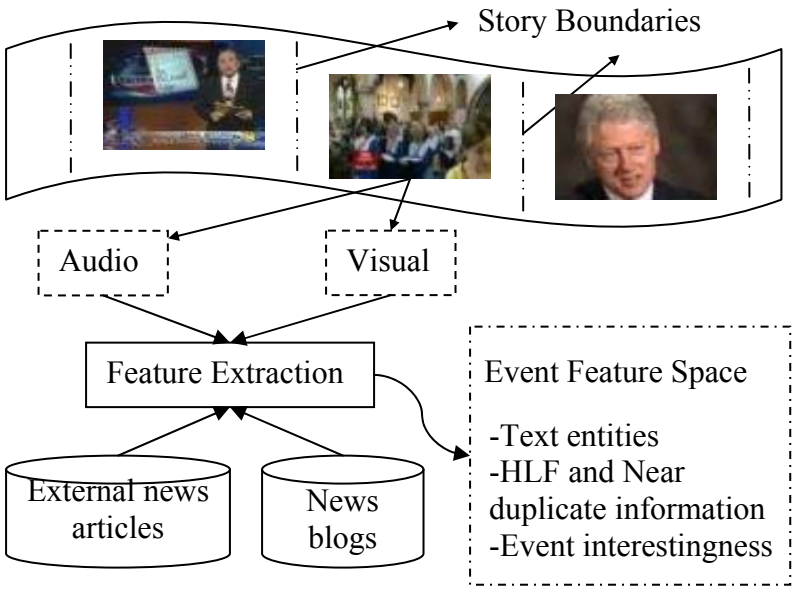


Figure 5.2 Extracting events entities from news video story

The event feature space we constructed at the story level consists of the following event entities: (a) text entities from speech; (b) HLF and near duplicate information; and (c) event interestingness from external sources. These features are raised from the video features extracted.

5. 2 Text Event Entities from Speech

Important aspects of an event like location and time are likely to be narrated by the speaker or reporter in news video. Thus, speech transcripts or closed captions contribute important sources to obtain such event entities. We choose to utilize speech transcripts and/or closed captions using the "bag-of-word" approach where the ordering of words within a sentence or phrase is disregarded. This assumption allows us to treat each news story as a holistic unit, enabling us to avoid issues such as the sentence boundary, sentence truncation and grammatical/word alignment errors and errors caused by machine translation. With the bag-

of-words that represents a news video story, the next step in the process is to identify and extract the various terms that belongs to different event entity type. The essential event entity types include: **Location**: {country, city, county, places of interest, etc}, **Time**: {video timestamp, or specific date mentioned, etc}, **Subject** {person's name, organization, etc}, **Object** {tangible like car, people, intangible like war, oil prices}, **Quantity** {numerical}, **Action** {death, birth, murder, etc} and **Descriptions** {other deeds}. In particular, this extraction process follows closely to the information extraction [Trec] or name entity extraction [Yang03b] processes in text processing.

The text retrieval community applies linguistic morphological analysis over the text documents to obtain the part-of-speech and subsequently employs a generic name entity extractor that uses a combination of syntactic and semantic pattern rules to automatically extract various named entities. Part-of-speech provides word level information such as which words are the nouns and verbs in the sentence. It is thus possible to apply syntactic patterns from part-of-speech to identify name entities (NEs). The performance of state-of-the-art systems [Trec, Voor04] for the detection of NEs for the names of person, organization and location can be as high as 95% for plain text documents. However, as speech transcripts are seldom perfect, it is thus not possible to carry out linguistic morphological analysis to obtain part-of-speech. We therefore rely only on semantic rules or keywords matching. Existing lists of semantic keyword lists that are available are location, time, subject (common names) and object. For actions, we further predefine a set of prevailing keywords that are descriptive of event's actions or activities in news. This list consists of approximately 1000 words which are automatically generated by gathering news snippets from the relevant period of the news video corpus (i.e. murder, killing, fire).

The location type of entity is seldom misrecognized or wrongly translated in the speech transcripts as they come from fixed vocabularies. In contrast, person names are far more vulnerable to errors as they are coming from non-vocabularies. An event usually has a distinct location. However, as there may be a number of locations mentioned in a single news video story, it is necessary to select the most representative one. This selection is done based on spatial relationship of location terms. For example, if "Iraq" and "Baghdad" are both mentioned in the story, "Baghdad" will be selected as the event location as Baghdad is the capital of Iraq. This rule is stemmed from the intuition that a more specific place mentioned in the speech transcripts tend to suggest that it is where the event occurs. The time entity of the event is taken by default to be the date of the news video when no other date information is found in the story. However, cue terms such as "yesterday", "two days ago" might signify that the event happens earlier.

To improve the accuracy of detecting person's names in news video, we gather text news articles of the period corresponding to the news video, and apply linguistic morphological analysis and generic named entity extraction on these articles. In the actual implementation, we use articles that are dated 3 days from a news video to obtain additional names. A supplementary name list is then dynamically generated for each date in the video corpus. As the news articles are grammatically correct, part-of-speech can be obtained to combine both syntactic and semantic parsing which allows a more complete detection of the person's name. Through this process, we are able to identify approximately another 40% of initially undetected names.

5.3 Visual Event Entities from High Level Feature and Near Duplicate Shots

Even though most event entities can only be obtained from speech, there are many audio and visual elements in the news video that can be important. For example: audio signatures like "engine noise" can indicate an aircraft taking off; "clapping or cheering" can indicate a large crowd; or visual scenes of "fire" can indicate events like fire outbreak or forest fire. This important information may not be available from text. For multimedia event, we are more interested in the use of high level features rather than low level features. The main reason is that the former offers more semantic information about the video content. For example, it is not semantically reasonable to assume a certain event like forest fire, tends to have more "red" pixels than "white" ones. We have discussed the use of 50 high level features in section 4 at the shot level. However, since a video event is based on news video story rather than shots, we need to appropriately model this inheritance relationship. Eqn 5.1 computes the confidence of a particular high level feature HLF_k arising from the shots within a news video story v.

$$Conf_{story}(v \mid HLF_k) = \max\{Conf_{shot}(s \mid HLF_k), s \in v\} \quad (5.1)$$

where $Conf_{shot}()$ is from Eqn 4.2. Here, we employ a greedy approach in which the confidence of the news video event for a particular HLF_k takes the max confidence value from the shots within the story. This is because if HLF "fire" is detected with high confidence in a particular shot, it can be conjectured that the story containing that shot should be associated to "fire" too.

Besides using high level features, near duplicates information is also employed. This is because we observe that reports of the same event can sometimes have the same footage or slightly different footage. For this, we employ D_s that is the matrix showing the near

duplication relationship from Eqn 4.1. The use of such information can provide linkage between news video stories which we will describe in detail in section 4.

5.4 Multimodal Event Entities from External Resources

Previous sections show how event related entities are derived from the video source. In this section, we present how related information from external news sources can be further leveraged into the event feature space. In particular, we introduce an "event-interestingness" modality, derived from news blog resources. Blogging has recently attracted much attention [Techno] as we see the increase in its internet utilization. There is a growing mass of people expressing their views and ideas on events happening around them in the form of web blogs. The events they commented on range from their everyday life, current news, animal rights issues, to rumors on celebrities. A typical web blog posting consists of text, images, videos and links etc related to a particular topic. Bloggers create the blogs with the idea of expressing themselves. But unknowingly, the materials they provided act as a source of unrefined knowledge, and recent web research [Trec, Wiki] found that valuable collective information (i.e. to support question answering) can actually be deduced from blogs.

"Event-interestingness" is the interestingness factor of a news event. For instance, how much attention is being received from the public for this news event? As discussed in Chapter 1, there is a mass of users who are constantly looking for interesting news information, and specifically to these searchers, returned results should be ranked accordingly to both relevancy and factors such as interestingness. In this work, we choose to rely on gathering statistics from "web activities" to obtain intrinsic information about the interestingness of certain news events. The rationale is that when an interesting event occurs,

the general public's attention will be focused on that event and web activities in the form of news commentaries and especially blogging will be high. We can observe a sharp rise in web activity for this event, "*the capture of Saddam Hussein*", which triggered a huge number of blog postings and news articles relating to him in December 2003. It was only in 2003 where worldwide blogging has just picked up. In addition, we can also see an overwhelming amount of web activities when Saddam Hussein was put to death in early 2007. Thus a sharp increase in postings on a topic usually suggests that an important event or interesting event has occurred. According to this phenomenon, we conjecture an implicit but direct correlation between web activities and the interesting events.

To measure web activities, we gather the statistics on web activities relating to a particular event. We first retrieve news blog postings relating to the news video events from Technorati.com [Techno], which is the largest online web blog search engine. The retrieval process is done using key event entities extracted such as subjects, location, time and actions. Given the list of relevant blog indexes, we proceed to count the number of relevant blog postings that are within a single-link away from the initial retrieval results. The reason for counting only single-link instead of multiple-link is that there are usually many external links (e.g. online commercial, link to other blogs etc) that may contain information unrelated to current topics. To further ensure that these postings are relevant to the particular news video story, two filtering criteria are employed. First, only postings which have overlapping name of *subjects (i.e. Saddam), location (i.e. Burma) or action (i.e. fire, flood)* in the blog-title will be considered. Second, the posting *date* must be around the same period as in the news video event time. An example plot of blog posting against time for "*Arafat*" in Nov 2004 is shown in Figure 5.3.

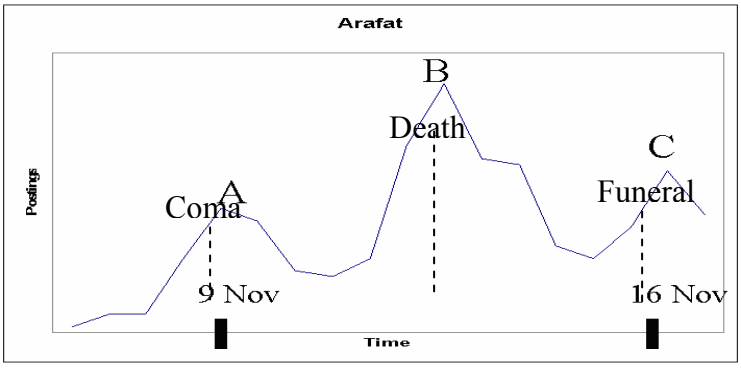


Figure 5.3 Blog statistics for "Arafat" in Nov 2004

From Figure 5.3, we can see a number of fluctuations in the number of postings. The total number of relevant postings collected for the entire period is about 5,000, and the difference between the days with the highest (304) and lowest (33) number of postings is close to 300. We found that the causes for the various peaks are: **(A)** when Arafat was reported to be critically ill and in coma in the hospital; **(B)** the highest peak occurs on the third day after Arafat is pronounced dead by the officials; and **(C)** reports and news on Arafat's funeral. From observations, we notice the number of blogs per day can fluctuate greatly across different periods of time. We therefore model event interestingness using a logarithmic function on the blog posting to smooth the trade-off as in Eqn 5.2.

$$Interest(Story_y) = \log(P_t) \qquad (5.2)$$

where P_t is the number of relevant postings in the given time period t (t is taken to be 1 day before and 2 days after the news event). To appropriately gauge a suitable time period, we carry out an experiment using a set of predefined topic terms to understand the distribution of blogs posting across time. From the results, we observe that this time-lap is usually one to two days after the occurrence of the major event. News often has a life-cycle after it has been published; it requires a time period for the public to get interested and start the discussion. After this period, it will begin to quiet down until the next wave of discussion

triggered by some new related events starts. However, it is also possible that a news event is predictable (i.e. elections) where intensive web activities are already seen days before the actual event.

5. 5 Employing Parallel News Articles for Clustering

The erroneous nature of speech transcripts makes it unlikely to obtain full aspects of an event. Furthermore, broadcast news is often presented in summarized form, and certain aspects of events might also be unavailable. This limitation prompts the use of relevant external resources, in particular, parallel text online news resources to supplement news video. We propose to leverage a combination of news articles and news video stories in the same clustering space. The rationale is to make use of the innate relationships between event entities from both sources of information to better induce the event clusters.

For example, considering the following two news videos (story1 and story2) of the same event but having missing entities due to speech transcription errors or machine translation. Text of story1: *"Chemical factory explosion kills more than ten people in California"*; Text of story2: *"Blast in Western United States … refinery death toll to 14"*. The text of story2 did not mention *California* and have many terms which are different to story1. In addition, we found that story2 actually have a higher similarity value to story3, which have terms like "*Death toll of the tsunami in Western Java rose to more than thousands, the red cross and the United Nations are ….*" reported in the same period. In this case, it is possible that story1 and story2 may be clustered wrongly into different clusters. It is important to provide a semantic bridge so that story1 and story2 can be clustered together. We leverage the news article of the "explosion event" to provide better and fuller

description like the text from the article: "*The number of death in the oil refining chemical factory located Western United States; California rose to more than 10. The blast was believed to be caused by ...*" We can see that text from this article overlaps with both story1 and story2. When employed together in clustering, this article causes the cluster centre of this "explosion event" to be shifted in a dimension closer to both story1 and story2, thus creating a higher probability of having story1 and story2 clustered in the same cluster. In this Section, we will describe how we obtain and extract entities from text articles to carry out clustering with the news video stories.

One of the most distinguishable differences between news video and other videos (documentaries, sitcoms, dramas, movies, etc) is that news video is usually a report of real-life happening around us. Due to limited air time, a news video cannot provide all the information but rather provide only, a summarized version of the event which contains enough information for general user appreciation. Intuitively, since news video stories are events which command high attention, there must be plenty of alternative sources of information which are directly or indirectly related to them. For instance, we can reasonably argue that video news is only a fraction of all reported news. In recent years, we observe that the growing trend of publishing online news, where news agencies are feeding news directly online to gather user hits. In fact, recent statistics [Sear] show more and more users are switching to obtaining news online, rather than on newspapers since this is the fastest medium for information exchange. News search engines such as Google News have already indexed news documents from over 4,500 websites to provide "instantaneous" news.

The efficient online information gathering platform facilitates the collection of external news resources related to news video. For news video based on recent news, we can

simply rely on "top stories" from existing news search engines on various news online sites as the news articles *links* are likely to be "alive". However, this technique will not work for news older than two weeks since the link is likely to be "dead". This can pose a problem as video corpuses are usually dated much earlier. To tackle this problem, we utilized the Highbeam Research [High] news archive database which archived over 400 online news sources. However, it is impractical to gather "all" news that falls within that period as the data can be overwhelming. In our implementation, we follow the line of thought of gathering top stories or at least articles which are highly viewed by searchers. The set of articles that we utilized: (a) consist of the top *n* articles each day that have the highest viewer rates (given as click rate); and (b) correspond to the same period as the news video.

5. 6 Temporal Partitions

In order to obtain the groupings of the instances in the video corpus, we devise an unsupervised clustering framework, *temporal multi-stage clustering* as shown in Figure 5.4. The multi-stage clustering framework uses text for the first stage of clustering and a combination of text and visual for the second stage clustering. The aim of first stage clustering is to identify events, and text is the best feature for event detection and tracking [Alla98]. As text features from speech transcripts are insufficient, we supplement it with the parallel news articles. This combination enables key events entities in events (from text) to be available for effective event clustering. The other reason for using only text in this stage of clustering is that it involves both video stories and parallel news articles (as parallel news articles do not have the associated visual features). Clustering at the first stage employs the hierarchical *k*-means clustering method based on text entities to obtain the initial clusters.

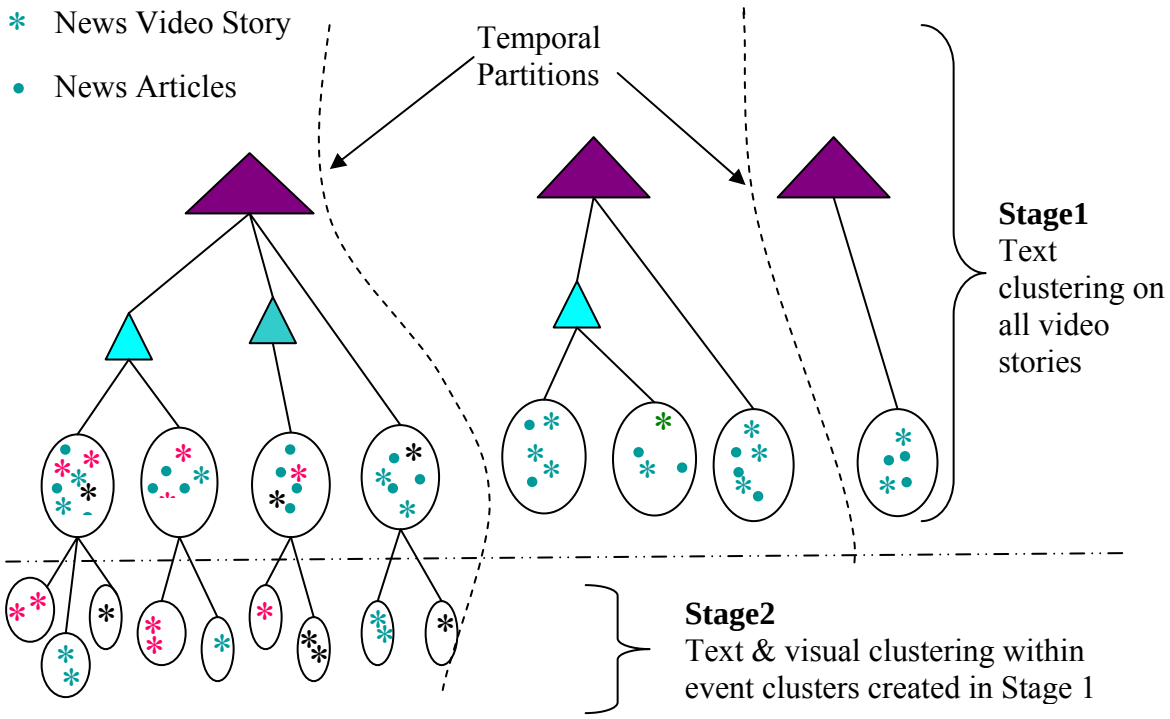


Figure 5.4 Temporal multi-stage event clustering

After the news video stories are clustered into event clusters, the next stage aims to further cluster each event into sub-events. It is observed that sub-events tend to involve non-text cues like scenes of live reporting on various aspects of the events. Such aspects are only identifiable using a combination of text and visual data frames. Hence, we employ a combination of visual features from news video stories along with the text entities at the second stage to refine the initial clusters. As clustering on a large scale is computationally expensive and large clusters tend to contain noise by outliers, we divide the video news stories into temporal partitions. The temporal partitioning aims to greatly reduce the computation time, and provides a smaller clustering space from which better clustering results can be obtained. To enable clusters from different partitions to be connected into an overall structure, we design the partitions with temporal overlaps. The following subsections

will provide the details of clustering and threading of the clusters in the temporal partitions.

5. 6. 1 Multi-stage Hierarchical Clustering

Clustering aims to partition the given set of data into subsets (clusters), such that the data in each subset (ideally) share some common traits - often proximity in a given feature space according to some distance measure. The *K*-means algorithm is commonly used to cluster instances based on attributes into *k* partitions. It is similar to the expectation-maximization [Demp77] algorithm for mixtures of Gaussians in that they both attempt to find the centers of natural clusters in the data. The idea of *k*-means is to define *k* centroids, one for each cluster. The next step is to take every point (instance) in a given dataset and associate it to the nearest centroid. After that, we will re-calculate *k* new centroids as centers of the clusters resulting from the previous step. The algorithm aims at minimizing an objective function, in this case the squared error function in Eqn 5.3.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \| x_i^{(j)} - c_j \| \tag{5.3}$$

where $\| x_i^{(j)} - c_j \|$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, it is an indicator of the distance of the *n* data points from their respective cluster centres. Hierarchical *k*-means clustering [Mlad98, Grob05] extends normal *k*-means by further performing *k*-means clustering within the clusters. This form of clustering can help to cluster instances into more *crisp* clusters as illustrated in the Figure 5.5.

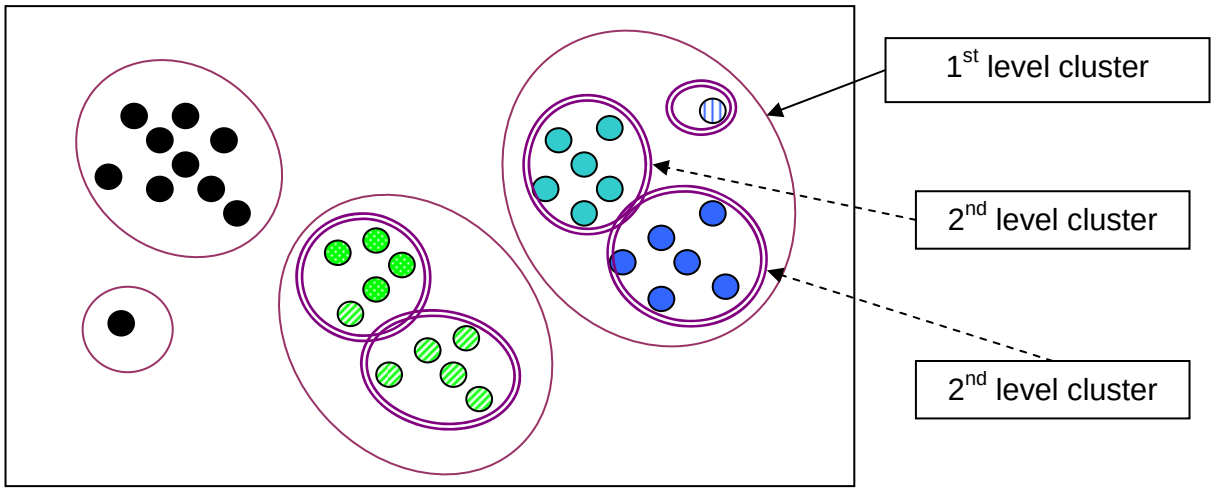


Figure 5.5 Hierarchical *k*-means clustering

One of the issues that *k*-means clustering faces is the selection of the value of *k* as this affects both clustering efficiency and effectiveness. In order to estimate a representative *k* value, we utilize a heuristic that is the *average number of news stories* covering an event. This number is estimated from a subset of the existing development data from TRECVID by calculating the number of news stories divided by number of distinct events. From experiments, this number was found to be 2.8 (approximately 3 stories per event). We preset $k = \lceil |V| / 2.8 \rceil$ (where V is the number of instances for the given dataset).

Stage 1 Clustering (text based): In text clustering, the documents are represented by sparse vectors of length equal to the number of unique words (or types) in the corpus. The components of each vector *v* have a value reflecting the occurrence of the corresponding word *w* in the document. A binary representation (with value of one or zero) can also be used to represent the presence or absence of the word. In our scenario, this will refer to if a particular location or subject is being mentioned in the news video or news article. We then employ the commonly used cosine similarity measure [Salt75] to computes the similarity between these two sparse vectors. We employ only event-related text entities such as the

location, person, time, etc extracted from the news video stories and the online text articles. The reason that we exclude non-event text is because they are shown to be less useful in discriminating events [Neo06b] and can even introduce noise in the process. Eqn 5.4 shows the cosine similarity measure:

$$Text\_Sim(v_i, v_j) = \frac{v_i . v_j}{|v_i| . |v_j|} \quad (5.4)$$

where $v_d = [w_{1,d}, w_{2,d}, ..., w_{t,d}]^T$, $w_{t,d} = tf.idf$. Here, *tf* refers to the frequency of word *w* in the document and *idf* is the inverse document frequency of *w* which indicates its rarity in the whole dataset. $Text\_Sim(v_i, v_j)$ provides a similarity value between 0 (no similarity) and 1 (v_i, v_j are identical). The clustering process then comprises the following steps:

1. *Choose k instances in the space. These points represent initial centroids.*
2. *Assign each instance to the cluster that has the closest centroid.*
3. *When all objects have been assigned, recalculate the positions of the k centroids.*
4. *Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.*
5. *Repeat Step 2 to 4 using a different set of initial centroids*

Figure 5.6 Algorithm for *k*-means clustering

As the algorithm is sensitive to the initial selected cluster centres, it is essential to perform the operation multiple times to obtain optimal clusters.

Stage 2 Clustering (text and visual combination): In this stage, we further perform clustering of the news video stories in each sub-cluster with the addition of HLFs and near duplicate information. This is reasonable as video news reports of a same or related event can have visually similar traits and video footages. Two additional measurements *HLF_Sim* and *ND* are fused together with *Text_Sim* as distance measures for clustering. *HLF_Sim* measures the similarity between the HLFs scores between the news video v_i and v_j based on

the Euclidean Norm on the 50 HLFs; and *ND* suggests if two video stories v_i and v_j contain any near duplicate shots (return 1 if there are near duplicates and 0 otherwise based on D_s, duplicate matrix from Eqn 4.1).

$$HLF\_Sim(v_i, v_j) = \frac{1}{Z}\sum_{k=1}^{50}(|\, Conf_{story}(v_i \mid HLF_k) - Conf_{story}(v_j \mid HLF_k)\,|) \qquad (5.5)$$

$$ND(v_i, v_j) = \max\{D_{s_k,s_l}, s_k \in v_i, s_l \in v_j\} \qquad (5.6)$$

$$Dist(v_i, v_j) = \alpha_1 \cdot Text\_Sim(v_i, v_j) + \alpha_2 \cdot HLF\_Sim(v_i, v_j) + \alpha_3 \cdot ND(v_i, v_j) \qquad (5.7)$$

where $Conf_{story}$ is the confidence score from Eqn 5.1 and Z is the normalizing factor. The final distance measure used for the second stage clustering *Dist* is given in Eqn 5.7 where α_1, α_2 and α_3 are empirically set to {0.5, 0.25, 0.25}. The clustering algorithm follows Step 1 to 4 shown in Figure 5.6.

Objective Function. One problem we observe with the traditional *k*-means clustering is that the clusters are formed based on achieving minimum total intra-cluster variance. The default objective function based on the squared error function in Eqn 5.3 can sometimes overlook the cluster density *(CD)* in big clusters, especially in datasets that contain plenty of outliers [Este97]. This phenomenon is imminent in our news video corpus where there are plenty of events that are reported only once. We therefore maximize *CD* (as the objective function) to ensure the quality of clusters by decreasing the chance of sporadic outliers being clustered together with major clusters. We define the cluster density *CD* for cluster C_j using Eqn 5.8:

$$CD = \frac{|\, CV_{c_j} \,|}{|\, C_j \,|} \qquad (5.8)$$

where $CV_{c_j} = \sum_{x \in C_j}(x - c_j)(x - c_j)^T$, $c_j = \frac{1}{|\, C_j \,|}\sum_{x \in C_j} x$.

CV_{c_j} is taken as a virtual representation of "cluster volume" [Bohm04] of cluster *j* and consequently, adding a sporadic outlier will dramatically increase its value.

5. 6. 2 Temporal Partitioning and Threading

Performing a single clustering on the entire news corpus [Neo06b] is straightforward and simple. However, we have mentioned earlier that such a clustering process has two major drawbacks. First, there is a large number of outlier news stories that are reported only once and this can lead to incoherent clusters. Second, the process is computationally expensive as the upper bound of *k*-means complexity [Har05] is $O(n^{dk})$ (*n* is the number of samples, *d* is the number of dimensions and *k* is the number of clusters). In order to tackle the two issues above, it is necessary to partition the data into suitable sizes for clustering. We use the temporal information (date) from the video to partition the news video corpus. This is because events are usually time dependent and news relating to the same event tend to be contained in the same period of time. However, the drawback of partitioning is that the overall structure may become disconnected. To overcome this problem and to preserve the context, we introduce an overlap between the temporal partitions so that overlapping event information may be used to link hierarchical structures induced in different partitions together. In our study, we empirically set the temporal partitions to five days with two days overlap. After obtaining the various hierarchical clusters for each temporal partition using the technique described in the previous Section, we need to link and thread these hierarchical structures together so that they can be used as an integrated structure during retrieval. This step involves determining similar clusters between different partitions so that they may be threaded together. To do this, we employ available event entities at cluster level. As the number of news video stories varies among clusters, it is necessary to extract

representative features from individual instances so as to represent the whole cluster effectively.

Extracting CRT. We define a cluster representative template (CRT) using the following fields: {**Text fields**: location, subject, time, action, description, object} {**Visual fields**: high level feature scores, near duplicate information}. For the selection of text entities at the cluster level, we take the union of all the text event entities (Section 5.2) from different news video stories and group them separately into the respective entity type (location, time, etc). Thereafter, we assign a confidence score to each of these entities based on their *tf.idf* scores. The formula for the confidence score of entity e_j is given in Eqn 5.9.

$$Conf_{text}(e_j \in C_k) = Freq(e_j \in C_k) * Log_2(\frac{M}{m}) \quad (5.9)$$

where e_j is in cluster C_k, M is the total number of instances in C_k and m is the number of instances containing e_j. Only event elements within each entity type with a confidence score above a pre-defined threshold δ_{type} will be added to the template. In our implementation, δ_{type} is controlled by the highest scoring entity $Conf_{text}(e_n)$ for a particular entity type l, and only e_j having scores near (taken as μ=0.9) to e_n are considered.

$$\delta_{type} = \mu \cdot \max(Conf_{text}(e_j)), e_j \in C_k, e_j \in type\ l \quad (5.10)$$

For the selection of visual entities, we use the HLFs and near duplicate information. The HLF representation of a cluster is obtained by averaging the particular HLF_l confidence (Eqn 5.1) scores across all news video stories contained in the cluster C_k, as:

$$Conf_{cluster}(C_k \mid HLF_l) = \frac{\sum Conf_{story}(v \mid HLF_l), v \in C_k}{\mid C_k \mid} \quad (5.11)$$

The near duplicate information is used to determine if two clusters have any near duplicate shots. This is found by D_s which is the matrix of similarity given in Eqn 4.1.

Matching CRT. With these cluster templates, we can then compare inter-cluster similarity and identify highly similar clusters. The function for comparing two cluster templates is:

$$\begin{aligned} Sim(CRT_k, CRT_l) = {} & \alpha_1(Text\_Sim(CRT_k, CRT_l)) + \\ & \alpha_2(HLF\_Sim(CRT_k, CRT_l)) + \\ & \alpha_3(ND(CRT_k, CRT_l)) + \\ & \alpha_4(cluster\_Sim(CRT_k, CRT_l)) \end{aligned} \tag{5.12}$$

where *Text_Sim*, *HLF_Sim* and *ND* is from Eqn 5.4, 5.5 and 5.6 respectively using entities found in Eqn 5.9, $cluster\_Sim = \frac{|C_k \cap C_l|}{|C_k \cup C_l|}$. *Cluster_Sim* indicates if two clusters contain similar elements. A high *Sim()* score will intuitively mean that the two clusters are closely related. Here, $\alpha_1, \alpha_2, \alpha_3$ and α_4 are set to {0.5, 0.2, 0.2, 0.1} empirically. Figure 5.7 shows the Topic Hierarchy that is the resultant structure from threading the partitions.

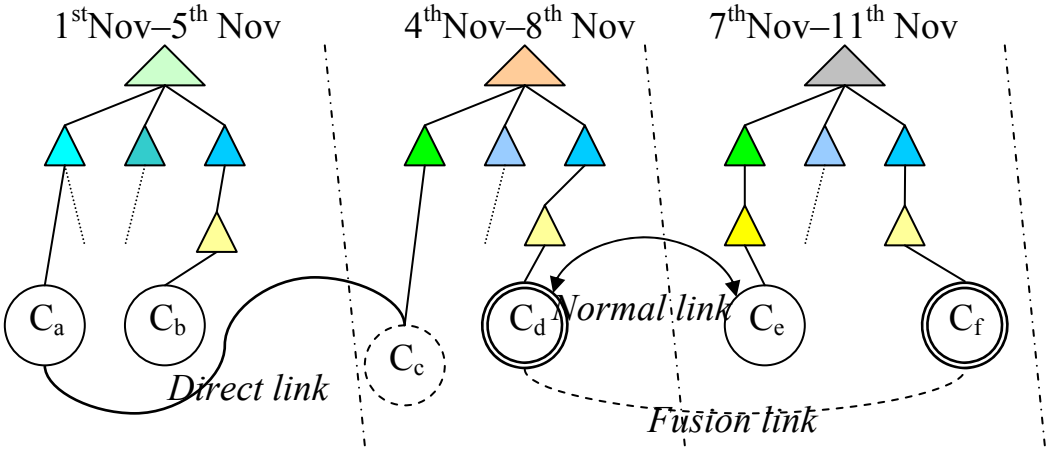


Figure 5.7 Threading clusters across temporal partitions in the Topic Hierarchy

To model the various linkage relationships appropriately, 3 types of links are defined: *"identical"*, *"near-duplicate"* and *"highly-similar"*. The first type, "identical", occurs when two clusters contain the same set of news video stories. This is possible due to the overlapping partitions. In this case, a *direct link* will be created between the 2 clusters.

The second type, "near-duplicate", occurs when *Sim()* score > δ_n (where $\delta_n = 1 - \alpha_4$). In this case, the two clusters must have overlapping news video instances and very high similarity. For this type of relation, we will create a *fusion link*. A fusion link allows us to upgrade the cluster by fusing both news video stories from both clusters together. Note that direct and fusion links will not occur within a single hierarchy since instances within clusters are *distinct*. The third type, "highly-similar" occurs when clusters have a similarity value of above the preset threshold δ_s. In this case, a *normal link* is created.

5. 7 Clustering Experiments

With the various techniques introduced above, it is important to understand how they can individually affect clustering performance. For evaluating the clustering performance, we manually go through the news video and perform topic based grouping. This is done using two weeks of news video data from both TRECVID 2005 and 2006 test datasets. The total duration of video time is 10 hours and 12 hours respectively. The experiments are designed to test the effectiveness of: (1) usage of parallel news articles for clustering; (2) clustering objective functions based on: (a) minimization of total intra-cluster variance; or (b) maximization of inter-cluster densities; and (3) usage of temporal partitions. The quality of a clustering is determined by analyzing the entire hierarchical structure. This is often done by using a measure that takes into account the overall set of clusters that are represented in the hierarchical tree. One such measure is the F_{Score} measure, employed by [Lars99, Zhao02] which is a modification of the usual F_1 measure.

We follow the evaluation as in [Lars99]. Given a particular class L_r of size n_r and a particular cluster S_i of size n_i, suppose n_{ri} documents in the cluster S_i belong to L_r, then the

F_{Score} of this class and cluster is defined as in Eqn 5.13 to 5.15.

$$F(L_r, S_i) = \frac{2 \cdot R(L_r, S_i) \cdot P(L_r, S_i)}{R(L_r, S_i) + P(L_r, S_i)} \tag{5.13}$$

$$F(L_r) = \max_{S_i \in T} F(L_r, S_i) \tag{5.14}$$

$$F_{Score} = \sum_{r=1}^{c} \frac{n_r}{n} F(L_r) \tag{5.15}$$

where R(L_r S_i) is the recall value defined as n_{ri}/n_r, and P(L_r S_i) is the precision value define as n_{ri}/n_i for the class L_r and cluster S_i. The F_{Score} of the class L_r is the maximum F_{Score} value attained at any node in the hierarchical clustering tree T as in Eqn 5.14. The F_{Score} of the entire clustering solution is then defined to be the sum of individual class F_{Score} weighted according to the class size as in Eqn 5.15, where c is the total number of classes. A perfect clustering solution will be the one in which every class has a corresponding cluster containing exactly the same documents in the tree, where the F_{Score} will be one. In general, the higher the F_{Score} values, the better the clustering solution is. The first series of runs are constructed as follows and the results are tabulated in Table 5.1.

baseline: (without parallel news, without temporal partitions)
B) baseline with maximizing inter-cluster densities objective function
T) baseline + temporal partitions
P) baseline + parallel news
BP) B + parallel news
BT) B + temporal partitions
PT) P + temporal partitions
BPT) B + parallel news + temporal partitions

Table 5.1 Performance of clustering for various runs with percentage in brackets indicating improvement over the baseline

T2005	**Baseline**	**B**	**T**	**P**	**BP**	**BT**	**PT**	**BPT**
F_{Score}	0.378	0.393(4)	0.434(15)	0.417(10)	0.437(15)	0.533(41)	0.499(32)	0.567(50)

T2006	**Baseline**	**B**	**T**	**P**	**BP**	**BT**	**PT**	**BPT**
F_{Score}	0.298	0.322(9)	0.388(30)	0.345(15)	0.375(26)	0.489(64)	0.455(52)	0.511(71)

From Table 5.1, we can draw the following conclusions. First, employing the correct objective function plays a significant role in improving the clustering performance. This can be seen in Run B, BP and BPT where they yield better results than their counterparts (Baseline, P and PT). In particular, significant improvement is observed in BPT over PT for both TRECVID corpuses. By analyzing the resulting clusters, we found that this is due to better modeling of the outliers as there are more clusters that only has one element.

Second, the use of parallel news is also effective as can be seen in improvements in clustering performance for P over B, and BP over B. In particular, the use of parallel news seems to be better on the corpus of TRECVID 2006 compared to TRECVID 2005. We conjecture that this could be due to the lower quality of machine translated text as the TRECVID 2006 video corpus contains a higher percentage of non-English videos. Thus TRECVID 2006 benefits more as compared to TRECVID 2005 from the use of parallel news in partially alleviating the errors and missing event entities. The lower quality of ASR text also provides one possible explanation on why the overall clustering performance in TRECVID 2006 is much lower than that in TRECVID 2005.

Third, the most significant improvement comes from the addition of temporal partitions that can be seen from run BP and BPT across both corpuses. This is mainly attributed to the nature of news video that is time dependent in nature. A significant difference in time usually means distinctive events. Thus partitioning the video corpus into

time segments is intuitive and effective.

The second series of tests access the improvement in performance on the second stage clustering. We would like to ascertain whether: (a) the use of visual features can supplement text features in clustering; and (b) which visual features are useful and to what extend. Table 5.2 tabulates the result from the second series of runs based on:

BPT) best run from first series of runs
BPT-H) BPT + high level features
BPT-N) BPT + near duplicate information
BPT-HN) BPT-H + near duplicate information

Table 5.2 Performance of clustering for second series of runs with percentage in brackets indicating improvement over the baseline

T2005	**BPT**	**BPT-H**	**BPT-N**	**BPT-HN**
F_{Score}	0.567	0.582(2.6)	0.602(6.7)	0.608 (7.2)

T2006	**BPT**	**BPT-H**	**BPT-N**	**BPT-HN**
F_{Score}	0.511	0.545(6.7)	0.546(6.8)	0.561(9.8)

From the results in Table 5.2, it is evident that visual feature are useful in further refining the clustering results arising from text features, as all three additional runs results in improvements in clustering performance. The best run comes from BPT-HN that yields an F_{Score} of 0.608 and 0.561 for TRECVID 2005 and 2006 dataset respectively.

In addition, we observe something peculiar about the results when comparing these 2 corpuses. The addition of high level features seems to be more effective for TRECVID 2006 dataset as compared to TRECVID 2005 even though the high level feature detection accuracies are lower (see Chapter 4.3). We speculate that this could be due to the following two reasons. The first reason could be that the overall clustering performance of TRECVID 2005 dataset is already high and further improvement is therefore limited. The second

reason could be the difference in dataset, as TRECVID 2005 consists of videos from 7 sources while TRECVID 2006 has 12, and TRECVID 2006 therefore has a high percentage of non-English news. This creates a paradigm for visual features to be more important when the text features are less useful.

Chapter 6

Query Analysis, Event Retrieval and Question Answering

Apart from having good features and sound understanding of news video, it is also necessary to interpret the user's intention to retrieve the most relevant segments given the user query. In this chapter, we discuss how query analysis is carried out with the help of external resources to enhance understanding and interpretation. In particular, we aim to extract query features such as (a) query-terms; (b) query high level features; and (c) query-class from the text query supplied by user. With the use of these query features, we can then perform precise retrieval at the story or shot level. Subsequently, we will present the integration of the event model into supporting various functions during retrieval. The process involves the discovery of a *query topic graph* that is a graphical structure containing relevant news video organized at an event level to the query. Differing from existing retrieval engines, the relevant news materials found through this query topic graph provide structured information that is useful in facilitating: (a) generalized topic viewing with key events viewing; and (b) question answering for context oriented or visual oriented queries.

6. 1 Query Terms with Expansion on Parallel News Corpus

Query terms (denote as q) provide the necessary initial context for retrieval. As the users' text queries are usually short and imprecise, it is necessary to carry out inference to

gather extra context, which is the basis for query expansion. Query expansion (the expanded query is denoted as *q'*) usually involves techniques such as: (a) finding synonyms of words; (b) finding the various morphological forms of words by stemming; (c) finding highly correlated words; and (d) re-weighting the terms in the original query. The goal of query expansion in this regard is by increasing the recall, while precision can potentially be increased as the returned result set will be more relevant or contains higher quality results. Without query expansion, many relevant items will not be retrievable even they may be semantically relevant.

As for video retrieval, query expansion is also important. Comparing to word documents, speech transcripts and closed caption texts are often imperfect. This effectively means that if a particular term that is critical for retrieval is "missing", there will be little chance of retrieving this piece of news. In order to perform query expansion effectively, we make use of a time dependent expansion using a parallel news corpus. We perform expansion by generating additional query terms using the following 2-step approach: (a) using the original query terms to retrieve relevant news articles from the parallel news resource, and (b) extracting terms from these news articles which have high mutual information [Kenn89] with the original query-terms. For example, queries like: "*Find stories related to flood*" or "*Find shots containing buildings covered in flood water*" will extract highly related terms like "*rain, flood, hurricane, yellow river* ...." Intuitively, mutual information measures the information shared between terms X and Y (i.e. knowing one of these variables how much will it reduce the uncertainty about the other). For example, if X and Y are independent, then knowing X does not give any information about Y and vice versa, so their mutual information is zero. At the other extreme, if X and Y are identical then

all information conveyed by X is shared with Y. As a result, the mutual information is the same as the uncertainty contained in Y (or X) alone, namely the entropy of Y (or X: clearly if X and Y are identical they have equal entropy). Formally, the mutual information of two discrete random variables X and Y can be defined as:

$$MI(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{6.1}$$

where *p(x,y)* is the joint probability distribution function of X and Y, and *p(x)* and *p(y)* are the marginal probability distribution functions of X and Y respectively. In our application, word probabilities, *p(x)* and *p(y)* are estimated by counting the number of observations of x and y in the total retrieved articles and normalizing by G, the size of the text article corpus. Joint probabilities, *p(x,y)* are estimated by counting the number of times that x is followed by y in a window of *w (w=5)* words and normalizing by G. Only the top *n* (n=10) terms that have the highest *MI* are added as the expanded query terms.

Query expansion especially in the domain of text question answering and retrieval rely mostly on documents obtained through search engines [Yang03]. This same idea can be applied to news video retrieval. As the online news search engines only index recent news articles, it may not be appropriate to use them for expansion. For example, assuming we are performing expansion to obtain additional words for *fire.* In the news video corpus, the only related articles to fire are the "forest fire" articles. However, through expansion using the recent news articles from Google, we might obtain irrelevant terms like "explosion, chemical, factory" because of the recent news on the explosion in a chemical factory. In this case, query expansion will cause the retrieved results to be noisier. To overcome this problem, we carry out query expansion using text articles that correspond to the same period as the video data (see Chapter 5).

6. 2 Query High-level-feature (HLF)

Relying retrieval solely on text terms from the query is not sufficient as it is not possible to model appropriately what the user may want. This is especially true if the user's queries are visually-oriented such as looking for specific scenes or objects in the news video rather than a designated piece of news. It is therefore important to have the ability to utilize video features such as HLFs to supplement the text features. In this thesis, we discuss the use of query-HLFs which measures the importance of a HLF to a query. If a particular HLF is deemed important to the query, it can intuitively mean that the existence of this HLF can greatly influence the relevance of the returned segments. One direct example is "*Find shots containing sky*".

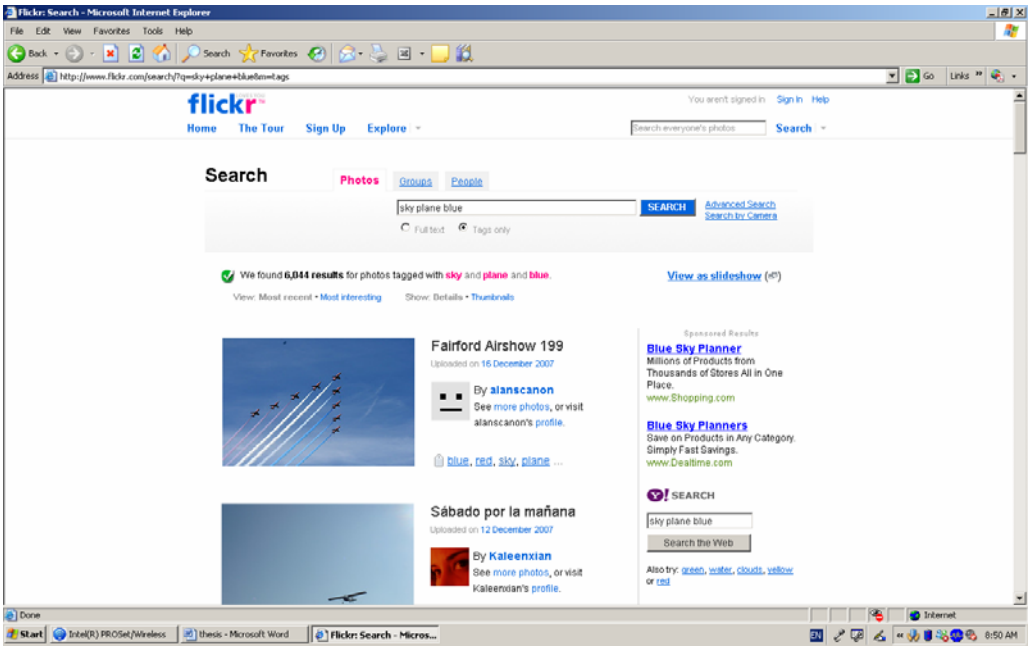


Figure 6.1 Retrieval from flickr using query "sky plane blue"

If "*sky*" happens to be an available HLF, the most appropriate retrieval method would be to locate shots which contain the "*sky*" HLF. Alternatively, there are also indirect examples such as "*Find shots containing a plane in air*". Note that the later query does not directly

mention "*sky*" but the HLF "*sky*" could be relevant. The current set of HLFs is limited and it is important to perform inference in order to support a wider range of queries. To induce the importance of a query to our set of HLFs, we employ both morphological analysis from WordNet [Fell98] and visual co-occurrence quality from Flickr [Flic].

WordNet has been a heavily utilized source of ontological lexical information in text retrieval. In text retrieval, systems usually relate terms lexically by means of hierarchical relationships [Mold01] such as synonymy, hypernymy and hyponymy. Instead of relying only on hierarchical relations, [Neo06] also chose to include lexical similarity measurement from the glosses, which contains the definition of the word. This is because additional properties describing lexical terms can sometimes be found from the gloss and this information can differ significantly from those extracted from a term's synonyms, hypernym and hyponym relationships. The definitions sometimes provide visual information about an object such as its shape, color, nature and texture; whereas the latter only provides direct relations (e.g., fire, blast: synonyms; airplane, bus: transportation mediums). However, the word "boat" can not be related to "water" by virtue of any relationship link in WordNet, but by its gloss – *"a small vessel for travel on water"*. It is therefore important to consider the definitional relationship to obtain better inference from the query to HLF. To employ this characteristic, we make use of the Resnik similarity [Resn99] metric in WordNet that equates similarity with the information content of the most specific common ancestor of the pair of words'as in Eqn 6.2.

$$\mathrm{Re}\,snik(w_i, w_j) = IC(lcs(w_i, w_j)) \tag{6.2}$$

where $lcs(w_i, w_j)$ is the most deeply nested concept in the *is-a* hierarchy that subsumes both w_i and w_j. *IC* calculates the *Information Content*, the common knowledge shared taken as the

negative log likelihood [Ross76]. Here, we factor in the expanded term weights from the previous step, and calculate the lexical similarity between all the expanded and original query terms to a HLF *k* using Eqn 6.3.

$$Sim\_Lex(Q, HLF_k) = \left(\frac{\sum_{w_q \in q'} \sum_{w_f \in HLF_k} Resnik(w_q, w_f)}{(|q'| \times |HLF_k|)} \right) \tag{6.3}$$

Even though lexical information can suggest a semantic relationship between the query and the high level feature, it may not necessarily imply that the suggested elements will actually be "*appearing*" together. After all, lexical similarity may not necessarily translate to visual co-occurrence. We know that a shot depicting a moving car tend to contain background with roads or tracks, while a shot containing a flying aircraft must be in the sky. However, such kind of intrinsic visual knowledge usually requires human intuition as there is no existing visual-object ontology which can bridge the gap.

To alleviate this problem, we tap onto the mutual information knowledge gained from large image repositories. The collective intelligence created by multiple users through the uploading of tagged images can sometimes help to predict visual co-occurrence quality. Flickr [Flic] is currently one of the largest image repository sites which contains millions of images. The images are manually tagged by the users who uploaded them and other interested parties who viewed them. The elementary use of tags is to facilitate searching using a text query. For example: a user looking for pictures of "*Bill Clinton*", "*US Flag*" or "*aircraft*" can first retrieve images with such tags and scroll through the list to choose those images they want. From the tags used for general retrieval, we attempt to leverage mutual association to infer the visual correlation of objects. The intuition behind this is that users tend to tag multiple meaningful words to an uploaded image and these words normally

correspond to a set of co-occurring objects or visual concepts. By studying the correlations between tags, it is possible to know whether certain types of objects have the natural tendency to be found together in the visual domain.

For example: in terms of lexical relations, the term "*plane*" is deemed to be related to both "*sky*" and "*train*". We first input the words separately: "plane", "train" and "sky" to carry out searching in Flickr and subsequently capture the number of relevant images found. This is done by extracting the total relevant images shown on the webpage. The operation is then performed using the bi-grams "plane sky", "plane train" and "train sky" to find images which contains both tags. The statistics obtained from Flickr is tabulated in Table 6.1.

Table 6.1 Statistics from Flickr using "Plane, Sky, Train"

Tags	Plane	Sky	Train	Plane & Train	Plane & Sky	Train & Sky
Total Found	**239,844**	765,784	1,291,346	3,325	25,788	10,024
Overlap "plane"	-	-	-	**1.5%**	**11%**	-

From Table 6.1, we can see that 11% of the images that contain "plane" also contain "sky", while in comparison, only 1.5% of such images contain "train". This visual co-occurrence quality tells us that the natural co-occurrence of "plane" to "sky" is much higher than that to "train". The appropriate way is to calculate the mutual information between the various bi-gram tags. However, as crucial information such as the total number of bi-gram tags as well as the number of unique tags is not available, we can only approximate the visual correlation; Eqn 6.4 shows the estimated visual correlation by using Flickr.

$$Sim\_Vis(Q, HLF_k) = \left(\frac{\max_{w \in q}(Flickr_{Count}(w \cap HLF_k)}{Flickr_{Count}(HLF_k)}\right) \tag{6.4}$$

where $Flickr_{Count}$ is the number of images with the relevant tags. In addition, we only make use of the original query terms instead of the expanded terms to minimize noise. Eqn 6.5

describes the final similarity measure from query to high level feature.

$$Sim\_HLF(Q, HLF_k) = \alpha \cdot Sim\_Lex(Q, HLF_k) + (1-\alpha) \cdot Sim\_Vis(Q, HLF_k) \quad (6.5)$$

with α empirically set at 0.4, determined through experimentation.

6. 3 Query Classification and Fusion Parameters Learning for Shot Retrieval

Query class is the next important query feature which aims to associate multimodal feature weights to query. The rationale behind the retrieval of a video shot can be different from a video story especially at the granularity of evidences and visual aspects. Most of the time, story retrieval is more contextual. For example "I want stories about Arafat's death", or "I want stories about the Iraq war", which are directed at the storyline of news video. However, queries like "*Find shots containing a car*" or "*Find shots of a computer screen*" are generally less dependent on text features. It is therefore necessary to know what features are important for what kind of queries. Query classification is frequently used in text question answering systems in TREC [Trec] as this knowledge will allow systems to apply appropriate retrieval schemes and even select possible answer targets. Given a virtually infinite number of queries, it is impractical to learn combination functions on a per query basis. A trade-off needs to be found between the difficulty of providing training data and the ability of capturing the idiosyncrasy of an individual query. As a compromise, a predefined set of query classes are preferred. The effectiveness of such query class dependent retrieval has been confirmed by our experiments on multimedia retrieval and many subsequent studies [Chua04, Yan04, Chua05, Huu05].

We adopt a query classification scheme that closely associates to the news video genre. The nine query classes are as follows: {PERSON, SPORTS, FINANCE, WEATHER, DISASTER, HEALTH, POLITICAL, MILITARY, GENERAL}. The GENERAL-class is

created to accommodate the queries that do not belong to any of the first eight classes. The main reason for this classification scheme is to create an explicit mapping of query class to type of news genre as it provides a good discrimination. For example: if the question is directed at sports, it is likely that the answers must come from sport news; and this is also true for financial news and weather news. These nine classes are also chosen because they cover over a wide range of queries. From our analysis of the datasets, the first eight classes cover more than 70% of the queries related to news video. In addition, these query-classes can be easily classified by using simple heuristic rules based on textual information alone. The firing rules for sports are terms like "sports, football, soccer, etc", that are related to sports. A list of queries from TRECVID search task and their classes is shown in Table 6.2.

Table 6.2 Examples of shot-based queries and their classes

Class	**Description/ Examples**
PERSON	looking for a person. For example: *"Find shots of Boris Yeltsin"*
SPORTS	looking for sports news scenes. For example: *"Score of basketball game"*
FINANCE	looking for financial related stories such as stocks, business Merger & Acquisitions etc.
WEATHER	looking for weather related shots. For example: *"weather for Washington DC tomorrow"*
DISASTER	looking for disaster related stories. For example: *"Latest update of death toll of train crash"*
HEALTH	looking for general health related stories
POLITICAL	looking for certain political happening, changes and congressman
MILITARY	involving military actions like war, missile attacks including terrorism
GENERAL	that do not belong to any of the above categories. For example: *"Find one or more people and one or more dogs walking together"*

Intuitively, queries from different classes will exhibit different characteristics and require different evidence to induce the answers. By simple observation, it may be possible to determine manually if a certain feature will be important. However, it can be difficult to obtain an optimal weighting manually if there are too many modalities to consider. For this,

we employ a Gaussian Mixture Model [Dasg99] approach to find the optimal feature weights of each class. To gather training data, we collect queries and their answer video clips (which is available from the TRECVID evaluation ground-truth) from past year's TRECVID (2003 and 2004) search task. The queries are classified into the eight classes as described. For those query classes that have less than 10 training queries, we manually create more training data for them. Using this training data, we can now estimate appropriate fusion parameters for each class.

The concept of parameter estimation for effective fusion of features is presented in many existing works [Yan06b]. In our implementation, we apply a straight forward method using a Gaussian Mixture Model (GMM) to find the density parameters using Expectation-maximization (EM) [Demp77]. In this case, we assume the following probabilistic model:

$$p(x \mid \Theta) = \sum_{i=1}^{M} \alpha_i p_i(x \mid \Theta) \qquad (6.6)$$

where the parameters are $\Theta = (\alpha_1, \ldots, \alpha_M, \theta_1, \ldots, \theta_M)$ such that $\sum \alpha_i = 1$ and each p_i is a density function parameterized by θ_i. In other words, we assume that we have M component densities mixed together with M mixing coefficients α_i and we are interested to determine the α_i and θ_i values as they form the required fusion parameters. The job of estimation is to devise appropriate parameters for the model functions we choose, with the connection to the data points being represented as their membership in the individual model distributions. We employ the EM algorithm to compute the parameters iteratively. We assume that each unobserved data $y_i \in 1,\ldots,M$ for each i, and $y_i = k$ if the i^{th} sample was generated by the k^{th} mixture component. If we assume that Y is a random vector, which takes a Gaussian model Θ^g, the probability density function can be simplified using Bayes's rule as:

$$p(y_i \mid x_i, \Theta^g) = \frac{\alpha_{yi}^g p_{yi}(x_i \mid \theta_{yi}^g)}{p(x_i \mid \Theta^g)} = \frac{\alpha_{yi}^g p_{yi}(x_i \mid \theta_{yi}^g)}{\sum_{k=1}^{M} \alpha_k^g p_k(x_i \mid \theta_k^g)} \tag{6.7}$$

The *d*-dimension Gaussian component distributions with mean μ and covariance matrix $\sum$, having $\theta = (\mu, \sum)$ is described in Eqn 6.8.

$$p_l(x \mid \mu_l, \sum\nolimits_l) = \frac{1}{\sqrt{(2\pi)^d \mid \sum\nolimits_l \mid}} \exp^{-\frac{1}{2}(x-\mu_l)^T \sum_l^{-1}(x-\mu_l)} \tag{6.8}$$

The mixing coefficients α_i are the means of the membership values over the data points and can be estimated using the maximization step as shown in Eqn 6.9 and 6.10. The component model parameters θ_i are also calculated by using data points that have been weighted using the membership values. For example, if θ is a mean μ.

$$\alpha_l^{new} = \frac{1}{N} \sum_{i=1}^{N} p(l \mid x_i, \Theta^g) \tag{6.9}$$

$$\mu_l^{new} = \frac{\sum_{i=1}^{N} x_i p(l \mid x_i, \Theta^g)}{\sum_{i=1}^{N} p(l \mid x_i, \Theta^g)} \tag{6.10}$$

With new estimates for a_i and the θ_i's, we go back to the expectation step to re-compute membership values. The procedure is repeated until the model parameters converge.

$$\sum\nolimits_l^{new} = \frac{\sum_{i=1}^{N} p(l \mid x_i, \Theta^g)(x_i - \mu_l^{new})(x_i - \mu_l^{new})^T}{\sum p(l \mid x_i, \Theta^g)} \tag{6.11}$$

In actual implementation, the training collection consists of $\{S_j, Class_j\}$, where S_j is the set of corresponding relevant shots for query class $Class_j$. A shot is represented by the corresponding set of features *x* comprising: (a) text features from speech at various level such as phrase and story; (b)HLFs; (c) image features; (d) near duplicate shot information; and (e) video captions from frames. Notice that the current set only contains five modalities

but the framework can be easily extended to more modalities when more modal features are available.

6. 4 Retrieval Framework

This Section illustrates the retrieval process which consists of three distinct stages: (1) story-level retrieval; (2) multimodal shot-level re-ranking; and (3) pseudo relevance feedback based on top return results; as shown in Figure 6.2.

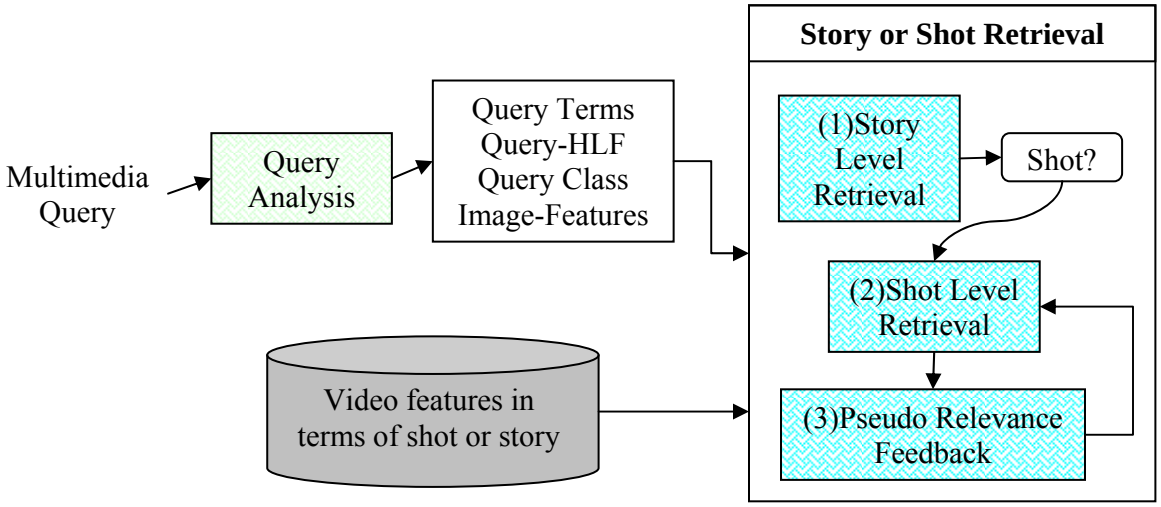


Figure 6.2 Retrieval framework

Story level retrieval. The first resolution in the framework deals with retrieval of a whole news story determined by the story boundaries as discussed in Chapter 4. Users searching for a specific piece of news usually have in mind what information they are looking for. They supply the queries usually in text and would like to view the list of relevant news stories. This type of retrieval is seen in many commercial news video retrieval systems such like Streamsage [Stre] and Blinkx [Blin]. However, these commercial systems mainly perform the retrieval by matching text features from speech transcripts to users' queries. In our system, we utilize two additional features discussed in Chapter 5 which are

the HLF and the cluster template. The retrieval function for a news video story is given in Eqn 6.12.

$$\begin{aligned} RF_{Story}(Q, v_i) &= \alpha_1 \cdot Text\_Sim(q, v_i) \\ &+ \alpha_2 \cdot \sum_{HLF_m \in v_i} \left[Conf_{story}(HLF_m \mid v_i) \times Sim\_HLF(Q, HLF_m) \right] \\ &+ \alpha_3 \cdot \max\left\{ Text\_Sim(q', CT_k \mid v_i \in C_k) \right\} \end{aligned} \tag{6.12}$$

where *Text_Sim()* is from Eqn 5.4, the second term computes the story level HLF score using Eqn 5.1 and Eqn 6.5, and the third term uses Eqn 5.4 to compute the text similarity between the *q'* and CT_k which is the cluster template of cluster C_k where $v_i \in C_k$. The parameters α_1, α_2 and α_3 are empirically set to {0.6, 0.2, 0.2}. Note that the scores of text entities are much higher than HLFs since story retrieval tends to invoke text features.

Shot Re-ranking. The next phase involves the retrieval at shot level. In our retrieval framework, we choose to further distinguish specific shots from story retrieval results if the user's query is specific to shots. Furthermore, available query-images or video shots which may be supplied by the user are also added as query features. With the query features, we can determine the relevance of shots by fusing the various multimodal features from video with the learned fusion parameters from the query classification described in Chapter 6.3. Eqn 6.13 shows the retrieval function for shots.

$$\begin{aligned} RF_{Shot}(Q, s_j) &= \alpha_c RF_{story}(Q, v_k \mid s_j \in v_k) \\ &+ \beta_c \cdot \sum_{HLF_m \in s_j} \left[Conf_{shot}(HLF_m) \times Sim\_HLF(Q, HLF_m) \right] \\ &+ \delta_c \cdot img\_sim(q_{images}, s_j) + \chi_c \cdot NDK(q_{images}, s_j) \\ &+ \gamma_c \cdot VCap(q, s_j) \end{aligned} \tag{6.13}$$

where $\alpha_c, \beta_c, \delta_c, \chi_c, \gamma_c$ are parameters for various query classes learned in Section 6.3; the first term makes use of the story retrieval function from Eqn 6.12, the second term computes

the HLF scores at shot level using Eqn 4.1 and Eqn 6.5, the third term measures the image similarity using the low level features, the fourth term checks if the query-image is a near duplicate of the sample shots, and the fifth term matches the query to video captions on the frame. Here we introduce *img_sim()*, a image matching function which computes a normalized Euclidean distance between the feature vector of low level features extracted from the available query-images and the shot key-frames. While *NDK()* indicates if the supplied query image or video shots are near duplicates, returning 1 if yes and 0 if no using D_s. Eqn 6.14 shows the V*Cap()* function.

$$VCap(Q, S_j) = \frac{1}{1 + \min\{MED(q, w), q \in Q, w \in VCap(s_j)\}} \tag{6.14}$$

Se Tom Deschle(0) Minority opoer

COMING UP Prbfit-Takin GM Works to beat W2K Bug Your

Figure 6.3 Video Captions (optical character recognition results)

The MED or Levenshtein distance [Leve66] between two strings is given by the minimum number of operations needed to transform one string into the other, where an operation is an insertion, deletion, or substitution of a single character. The score of *VCap* is taken as the inverse of the MED. This means a video with exact correspondence will yield a score of 1.

Pseudo Relevance Feedback. After getting the list of relevant shots, the next stage will be the pseudo relevance feedback (PRF). The positive use of PRF to re-adjust fusion parameters have been seen in [Chua04, Neo06]. The main reason is that the initial fusion

parameters are introduced based on the query class. This may not be good enough at an individual query level. The relevance feedback model proposed by Rocchio [Rocc71] modifies the query vector by iteratively increasing the weights of terms contained in positive documents and penalizing the terms in negative documents. However, since our process is PRF, getting positive instances is a challenge. For this, we make use of the top *n* returned shots (denoted as top_n, *n=15*) as positive samples for feedback. The rationale is to refine the fusion parameters directly using these top results so that they can better model the query. Using these top returned results, we re-train the fusion parameters following the EM process in Chapter 6.3. A set of new fusion parameters $\alpha_r, \beta_r, \delta_r, \chi_r, \gamma_r$ is discovered for query *Q*. The PRF retrieval function for shots is given in 6.15.

$$\begin{aligned} R\_RF_{Shot}(Q, s_j) = {} & \alpha_r R\_RF_{story}(Q, v_k \mid s_j \in v_k) \\ & + \beta_r \cdot \sum_{HLF_m \in s_j} \left[Conf_{shot}(HLF_m) \times R\_Sim\_HLF(Q_r, HLF_m) \right] \\ & + \delta_r \cdot img\_sim(q_{image}, s_j) + \chi_q \cdot NDK(q_{images} + top_n, s_j) \\ & + \gamma_r \cdot VCap(Q, s_j) \end{aligned} \tag{6.15}$$

Besides employing PRF for re-training of fusion parameters, we can also make use of the cluster information to propagate scores to stories that are contained in the same cluster C_p as the top returned shots. This is reasonable since stories contained in a cluster are effectively both contextually and visually similar. The story retrieval function with PRF is shown in Eqn 6.16.

$$\begin{aligned} R\_RF_{Story}(Q, v_i) = {} & \beta_1 \cdot Text\_Sim(q', v_i) + \\ & + \beta_2 \cdot \sum_{HLF_m \in v_j} \left[Conf_{story}(HLF_m \mid s_j) \times R\_Sim\_HLF(Q, HLF_m) \right] \\ & + \beta_3 \cdot \max\left\{ Text\_Sim(q', CT_p \mid v_i \in C_p) \right\} \\ & + \beta_4 \cdot \left\{ \frac{\sum_j 1,\ j \in C_p \wedge j \in top_n}{n} \right\} \end{aligned} \tag{6.16}$$

The newly introduced term computes the number of top retrieved shot in the same cluster as the story. The weights are re-distribution over the parameters $\beta_1, \beta_2, \beta_3$ and β_4 in the proportion of {0.4, 0.2, 0.2, 0.2} to accommodate for the new cluster score.

The next entity to benefit from PRF is the query HLF. Treating the top returned shot as "*correct samples*", we can adjust the query HLF by increasing the importance of HLF that is prominently present. For example: if 10 out of 15 top returned shots are deemed to contain HLF "*sky*" with a high confidence, then it can indicate "*sky*" is an important feature for that particular query.

$$\begin{aligned} R\_Sim\_HLF(Q, HLF_k) = {} & \beta_1 \cdot Sim\_Lex(q', HLF_k) \\ & + \beta_2 \cdot Sim\_Vis(q, HLF_k) \\ & + \beta_3 \cdot \left\{ \frac{\sum_{s_j \in top_n} Conf(s_j \mid HLF_k)}{n} \right\} \end{aligned} \tag{6.17}$$

We add a third term that accounts for the average detection confidence of a particular HLF in the top *n* shots to Eqn 6.5, resulting in Eqn 6.17. Similarity, the weights are re-distributed, in the proportion of {0.4, 0.3, 0.3} to accommodate for the term.

Finally, we add the key-frames of the top *n* returned shots to the existing pool of q_{image}. During re-ranking, shots that are deemed to be duplicates or near duplicates can be ranked higher. Through this PRF process, the retrieval is more specific and customized for the query.

6. 5 Browsing Events with a Query Topic Graph

A query topic graph is defined as a graphical interlinked structure containing news event materials (news video and articles) relevant to a query. In most video retrieval systems like Blinkx [Blin] and Streamsage [Stre], the relevant materials are usually retrieved as a

list-like structure with indications of their relevance to the query. Given a query from the user, our system first performs retrieval to obtain the relevant video stories and/or articles that contains the query terms. These retrieved documents are then mapped to the respective clusters containing them in the topic hierarchy as shown in Figure 6.4. The resultant sub-graph is the query topic graph which consists of a set of clusters as well as links related to the query.

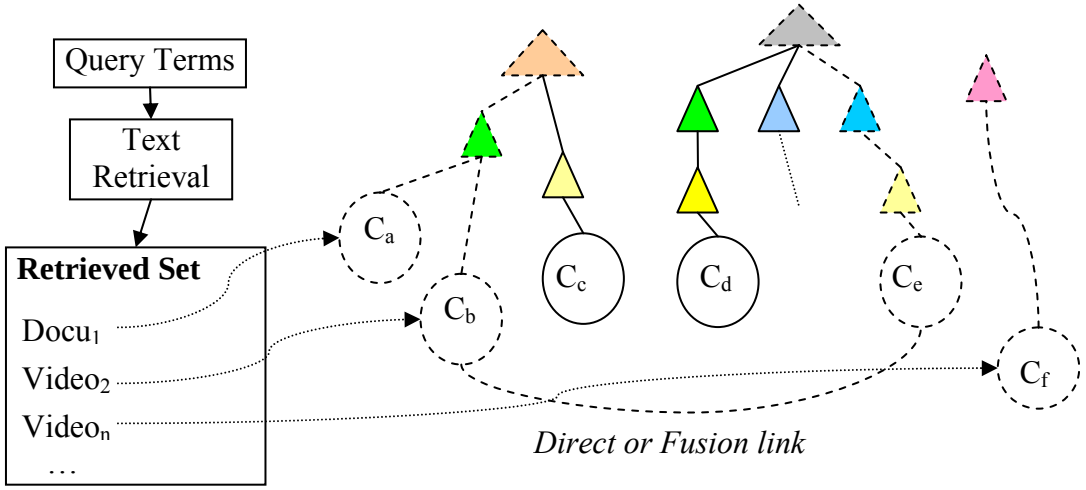


Figure 6.4 Query topic graph (denote by dashed lines)

To leverage on the links in the topic hierarchy, we further identify clusters that are children of marked clusters or can be reached by a single direct or fusion link through the use of Topic Hierarchy found from Chapter 5. This can effectively lead us into events or topics which might be initially un-retrievable using the supplied query. We introduce two browsing methodologies on top of the traditional linear rank list browsing: (a) hierarchical relevancy browsing which allows the user to view the news events in an interlinked structure based on the query topic graph; and (b) overview browsing on the topic in an evolutionary manner. We now describe how the two browsing methodologies can be achieved and performed.

Hierarchical Relevancy Browsing. From the query topic graph, we obtain the interlink structure as shown in Figure 6.5 by removing redundant nodes. We then make use of the story retrieval function from Eqn 6.12 to provide the score of the news video. For illustration, news videos that are highlighted by red borders have the highest score.

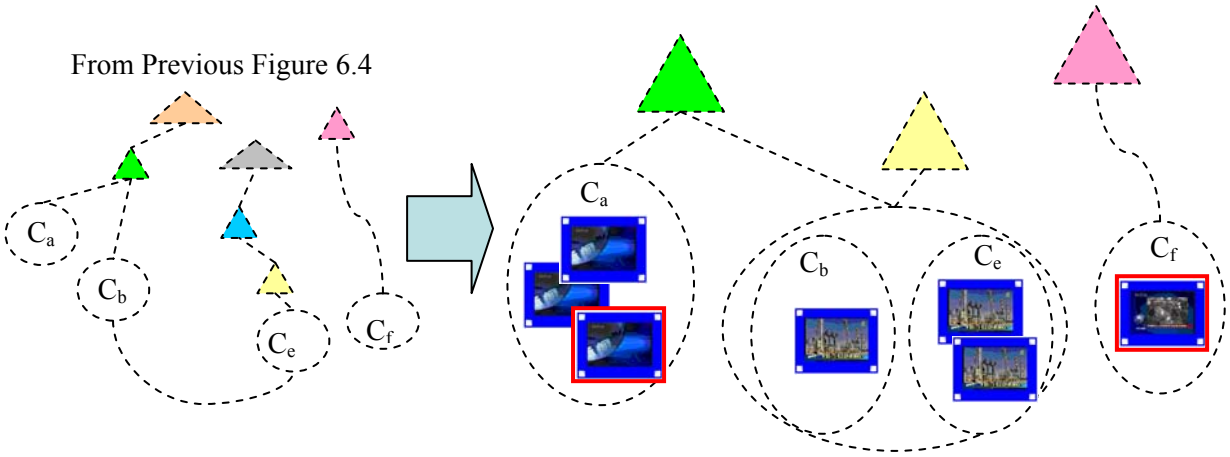


Figure 6.5 Interlinked structures from query topic graph

During hierarchical relevance browsing as shown in Figure 6.6, the user will be first presented with the top scoring news videos as shown on the left. The user can then "traverse" the interlink structure by: (a) click on "Show Cluster": viewing the other news videos in the same cluster; (b) click on "Up": moving up the hierarchy to display the list of clusters available at its parent node; or (c) click on a cluster: selecting the cluster when at a parent node.

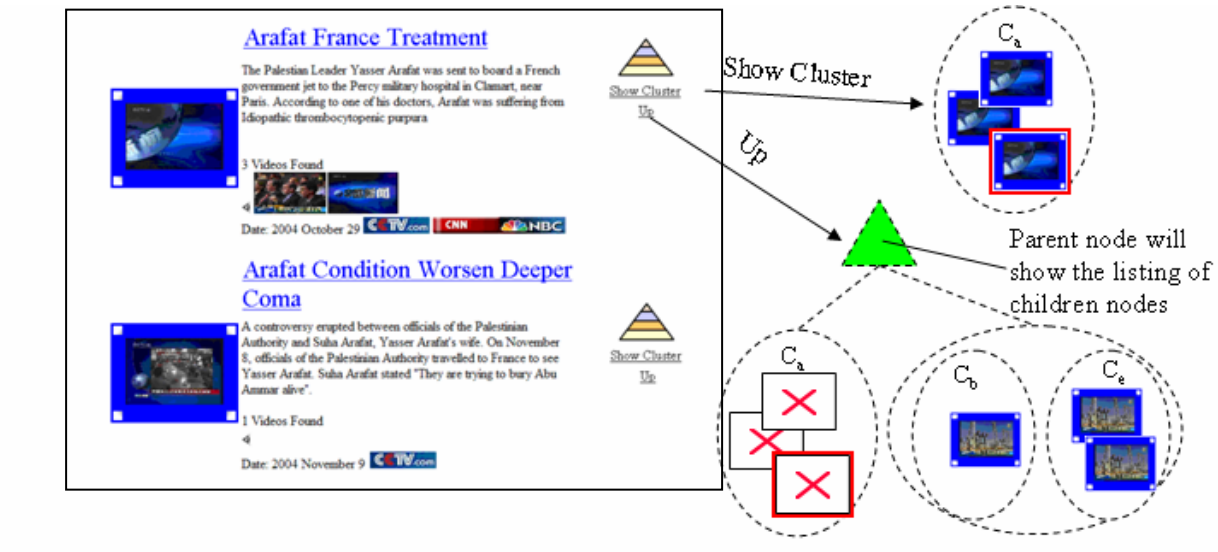


Figure 6.6 Hierarchical relevancy browsing using interlinked structures

Interestingness Browsing (Topic Evolution) Browsing. This second browsing technique leverages the interestingness facet of an event. We have discussed earlier that the interestingness of an event can be used to measure its importance to the topic. By viewing the events which are most important or interesting to a topic, we can grasp a good overview for that topic. For this, we integrate the interestingness factor from Eqn 5.2 into the story retrieval function from Eqn 6.12.

$$IRF_{Story}(Q, v_i) = \alpha \cdot RF_{story}(Q, v_i) + (1-\alpha) \cdot Interest(v_i) \qquad (6.18)$$

The new equation can be varied to give weights that bias either toward relevancy or interestingness. The following Figure 6.7 shows the results for "Arafat" when α is set to 0.6.

Figure 6.7 Topic evolution browsing for "Arafat" in Oct/Nov 2004

Notice that the top 3 results when arranged in chronological order show the 3 key stages in the search topic "*Arafat*" for that specific period of time. This browsing mode allows viewing of the most prominent event or events arising due to major changes. Viewing topic evolution will require news videos to be sorted in chronological order. It is clear that a simple linear rank list will not be suitable since the top result would always be one that has the earliest date. We therefore introduce a "full/half display" in which news videos that have higher score are displayed larger and with more details.

News videos with lower scores are displayed smaller as seen in the left-hand side of Figure 6.7 (5th Nov-Palestinian Authorities Arafat Brain-dead, 8th Nov – Palestinian

Authorities Travel France Arafat, 10th Nov – High Religious Dignitary Visited Arafat). This topic evolution browsing mode allows users to flexibly specify the period (starting and ending date) which he is interested in, and also dynamically change the value of α so as to view results tuned more towards relevancy or interestingness. The algorithm for displaying the videos is given in Figure 6.8.

1. **Ranking**: For selected period t {
    a. Obtain the corresponding query topic graph
    b. Apply α from user
    c. Rank news video v_i in topic query graph using Eqn 6.1 }

2. **Displaying**: For all v_i in query topic graph {
    //in chronological order
    a. if score $(v_i) > \phi_{full}$ then *display_full*(v_i)
    b. if score $(v_i) > \phi_{half}$ then *display_half*(v_i) }

Figure 6.8 Algorithm for displaying topic evolution

For both browsing methods, we display a set of keywords describing the news video to allow the user to grasp the content quicker. This is done by selecting text entities that have a *tf.idf* value following Eqn 5.4. This descriptor is limited to 12 words.

6. 6 Context Oriented Question Answering

While the topic evolution browsing and hierarchical browsing based on story retrieval provides the overview of a topic at the global level, question answering aims to return precise answers to specific questions posed over the news topic. The users expect the system to return short video segments with speech containing the answers. Applying the question answering to news would explicitly mean returning the relevant aspects of an event as answers. It can be seen in a personalized retrieval setting in which a user may further

request for **details** of different aspects of news such as "*How old was Arafat when he passed away?"* (looking for a number relating to age) or *"When did Arafat pass away?"* (looking for a date). This is similar to the well-known TREC Question Answering task. The task starts by defining a topic, for example "Crip", and then follows by a series of questions relating to the topic such as: When was the first Crip gang started? Which cities have Crip gangs? What is their gang color? This information seeking process is reasonable as users tend to have queries centering around a particular topic. For this task, getting the initial set of documents or materials that are relevant to the topic is critical and it is therefore necessary to leverage the use of the query topic graph. In this section, we will describe how we apply existing question answering technology together with the query topic graph to support question answering.

6. 6. 1 Query Analysis for Answer Typing

Most existing text-based question answering systems in TREC [Trec] rely on the inferred *answer-type* (whether the query is looking for a name, time, location, etc) as a guide when choosing possible answer candidates. This answer-typing is similar to query classification but is more complex since this is more fine-grained. We make use of an existing answer-type list from [Yang03] which contains the following answer types: {HUM_BASIC: person name, person title, HUM_ORG: organization name, LOC_BASIC: Country, State, TME_: date, year, month, time, Num_: age, count OBJ_: aspects of objects like color, full list in Appendix I}. This list is used for answering open-domain questions which can come from any information source. As our domain is news, we refine the list of answer-types to those which are relevant to news or events. A sample of queries, their firing rules and the answer-types are shown in Table 6.3. The topic of the query is taken as the

named entities, nouns or noun phrases mentioned in the query obtained through parsing.

Table 6.3 Sample queries with their answer-types

Topic	Question	Firing Rule	Answer-type
serial killer	What is the name of the serial killer?	"name"+"person"	HUM_PERSON
tornadoes	Which are the states which suffered tornadoes?	"Which"+"state"	LOC_STATE
Saddam Hussein	When was Saddam hang?	"When"	TME_DATE
Osama bin Laden	Where is Osama?	"Where"	LOC_BASIC
Olympic	Where is Olympics held in 2008?	"Where"	LOC_BASIC
Olympic	Which country is hosting Olympics 2008?	"Which"+"country"	LOC_COUNTRY
Olympic	Which city is hosting Olympics 2008?	"Which"+"city"	LOC_CITY

The issued query can contain a number of useful information sources which can help suggest relevancy. For example: the last three sample queries in Table 6.3 are concerned with the location of the Olympics 2008. From the way the query is asked, we can conjecture that it is directed at finding a generalized location, or a country or a city. Intuitively, if the answer-type is "LOC_BASIC", it can be answered with any entities from "LOC_COUNTRY", "LOC_CITY", or simply any entities which are location tagged. However, if the query is specific like in this case like, "*Which city is hosting Olympics 2008?*" returning "China" instead of "Beijing" will not be correct.

6. 6. 2 Query Topic Graph for Ranking

With the knowledge of answer-type, the system can now effectively narrow down the search range by looking for the presence of such candidates. However, it is still necessary to have an initial set of retrieved documents answer selection can be based on. We

therefore make use of the query topic graph obtained through the process mentioned in the previous section using the supplied topic. To obtain the best answer candidate, we employ the density-based ranking [Lee01] which uses a minimal distance between the matched words to measure the proximity of locality context. This is rationale as answers have a tendency to appear near to query terms due to linguistic coherence [Voor04]. For example: *"How many people were injured in the accident?"* It is not hard to imagine that the sentence which can contain the required answer is likely to contain "accident" or "injure" or "people" or something close to the above mentioned.

We apply this concept to measure how "dense" an answer candidate is by the overlap between: (a) query terms to words around the answer candidate in the phrase level, and (b) query terms to cluster template entities in the query topic graph. The scoring function is given in Eqn 6.19.

$$SF(e \in v_k) = \sum_i \alpha_i \cdot F^j \qquad (6.19)$$

where $\sum \alpha_i = 1$. The various features F used are: F^0 is the story retrieval score using Eqn 6.12, F^1 when the predicted answer genre matches the named entities genre (1 if there is a match and 0 otherwise); F^2 for word answer density (% overlap and distance between expanded query terms Q' and phrase containing the answer); and F^3 for cluster ranking density (% overlap and distance between expanded query terms Q' and event terms in the cluster template in the query topic graph).

6. 6. 3 Displaying Video Answers

The news video with the highest scoring answer will be returned as in Fig 6.9 & 6.10.

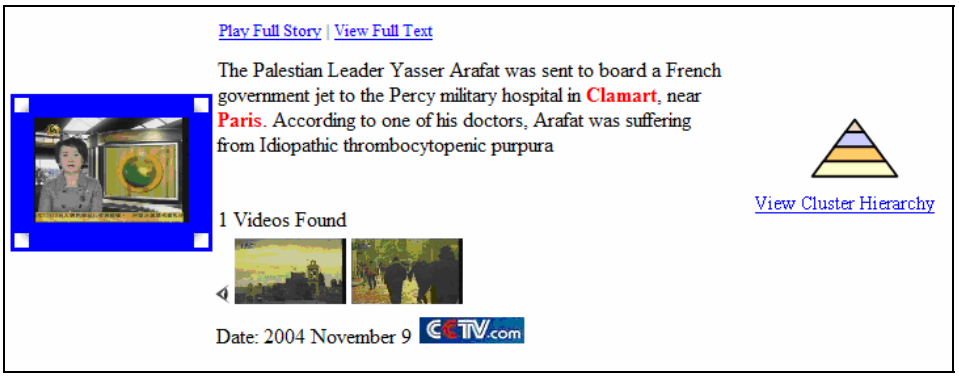


Figure 6.9 Result of "Where was Arafat taken for treatment?" (answers in red)

Figure 6.10 Result of "Which are the candidate cities competing for Olympic 2012?"

The news video is displayed together with the segment of speech which contains the predicted answer. For easy viewing, partial text containing the answer from the speech transcript is displayed along with the results as shown in Figure 6.9 and 6.10. Promising candidates are automatically highlighted in red.

6. 7 Visual Oriented Question Answering

Beside the context oriented questions which deal mainly with text, users may also be interested in visual-oriented questions. Examples like "*Find shots or scenes containing*

Yasser Arafat" or "*Find scenes of people holding signs or banner*". These visual questions would require segments of videos containing the required visual elements as evidence. Unlike questions directed at a factual aspect where a short segment of news video containing narrated speech is returned as the answer, the shot is chosen as a segment for visual-oriented questions. This is because shot is currently the smallest addressable video unit meaningful to users as defined in TRECVID [Trecvid]. The major difference between contextual question answering and visual question answering is the satisfaction of searching. While most users are content with one correct contextual answer, they would usually be interested to see more visual answers or video-footage. This is because footage containing what they want to see can be different across news stations due to video broadcasting rights and other exclusive rights issues.

For locating answer shots to visual oriented questions, we aim to provide as many relevant video segments as possible for the user. We make use of the query topic graph to provide semantic event cluster information as similar context events tend to have similar footages. To ensure higher recall, we expand the original query topic graph by considering immediate clusters which can be reached by "normal" links as seen in Figure 6.11.

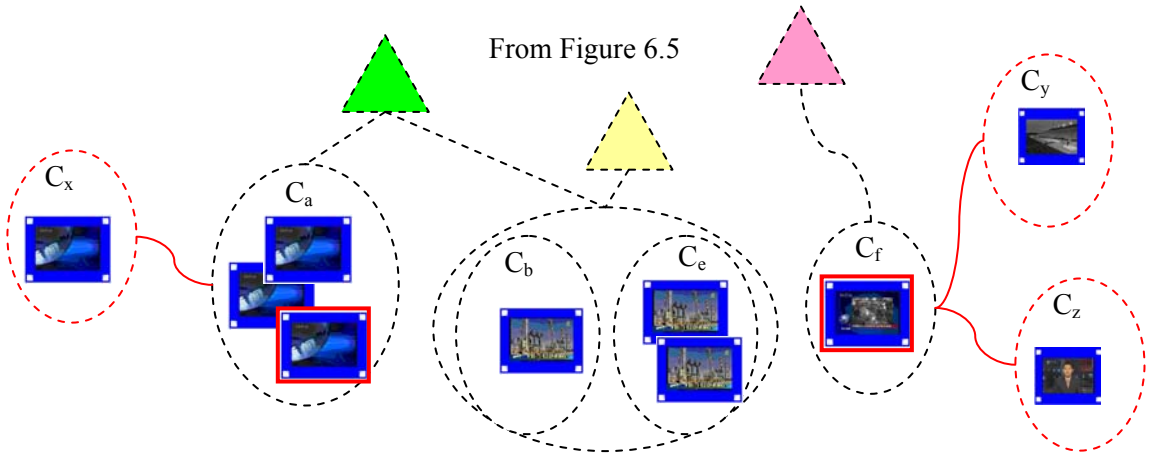


Figure 6.11 Expanded query topic graph (expanded portions denote by redlines)

The news videos in the expanded query topic graph are then used as the set for finding the answer candidates using the retrieval framework for retrieving shots in Section 6.4. The scoring function is given in Eqn 6.20. Figure 6.12 shows the results of a visual oriented query.

$$SF_{shot}(Q, s_j) = RF_{Shot}(Q, s_j), s_j \in Expanded\_topic\_graph \quad (6.20)$$

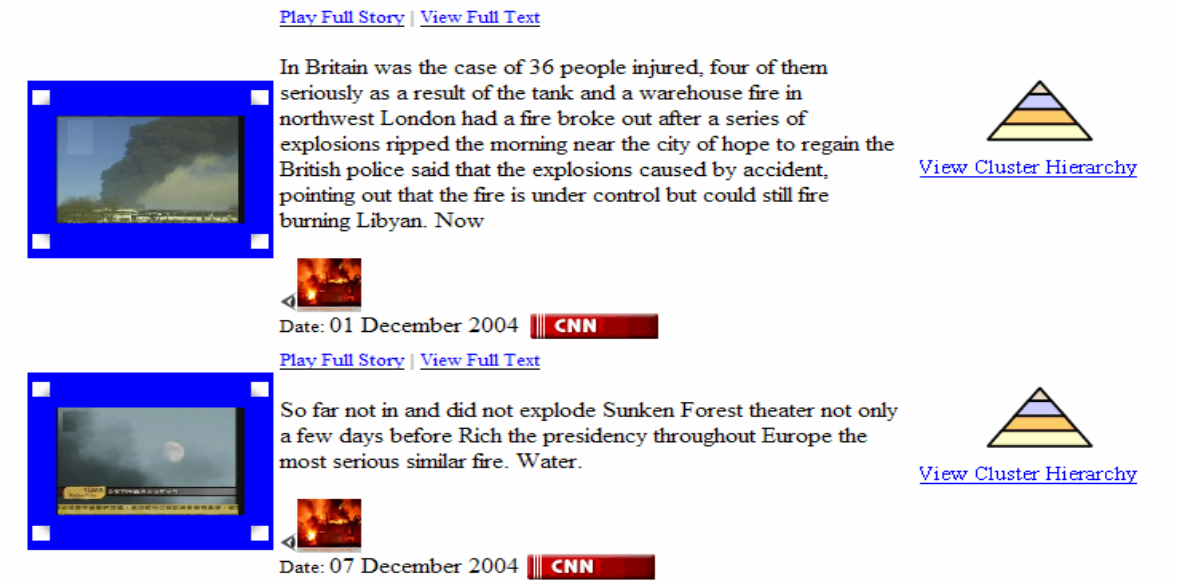


Figure 6.12 Result of "*Find shots containing fire or explosion*?"

Chapter 7

Retrieval Experiments

With the various techniques introduced and discussed in Chapter 6, it is important to understand how they individually contribute to the overall retrieval performance. Three sets of experiments are designed in this chapter to conduct comprehensive tests to test the performance of various key techniques. For comparison purposes, we follow closely with the specifications given in the TRECVID automated search task (details are given in Section 7.1). The first set of experiments is designed to test the multi-resolution retrieval framework introduced in Chapter 6, in particular, the use of query classification, external parallel text corpus for query expansion, high level features, and pseudo relevance feedback. The second set of experiments provides a user-based study to assess the performance of various browsing techniques using the event query topic graph. The third set of experiments focuses on question answering accuracy.

7. 1 Experimental Setup for TRECVID

The evaluation methodology follows the standard evaluation procedures carried out in TRECVID [Trecvid]. The search topics are designed as multimedia descriptions of an information need, which might contain not only text keywords but also possibly video, audio and image samples. Typically, the topics include requests for some specific items, specific

people, specific facts, instances of categories and instances of activities. In analogy to "document" in text retrieval, TRECVID adopts the basic video units to be retrieved as video shots, which is defined as a single continuous camera operation without an editor's cut, fade or dissolve. In particular, the TRECVID 2005 and 2006 datasets are used for evaluation in this thesis.

TRECVID2005: The video collection includes a 170-hour (nearly 150k shots) multilingual news video captured from MSNBC (English), NBC Nightly News (English), CNN (English), LBC(Arabic), CCTV(Chinese) and NTDTV (Chinese) in late 2004. The corpus is split into a development set including 74,532 shots representing 80 hours of video and a search set including 77,979 shots representing another 80 hours of video.

TRECVID2006: The video collection consist of 160-hour multilingual news video captured from NBC Nightly News (English), CNN LiveFrom (English), CNN NewsLive (English), MSN (COOPER), LBC LBCNAHAR (Arabic), LBC NEWS (Arabic), LBC HURRA (Arabic), CCTV DAILYNEWS (Chinese), PHOENIX GOODMORNINGCN (Chinese), NTDTV FOCUSINT (Chinese) and NTDTV ECONFRNT (Chinese) in late 2005. The total number of shots is 79,484, which is significantly less than TRECVID 2005 due to the fact that the average shot time is longer. Unlike the previous year, this set of data is set for testing purposes as the development set reuses the whole of the TRECVID 2005 corpus.

The search task comprises three types as shown in Figure 7.1. In particular, we follow the specification of the automated search task, which requires systems to carry out retrieval without any form of user intervention.

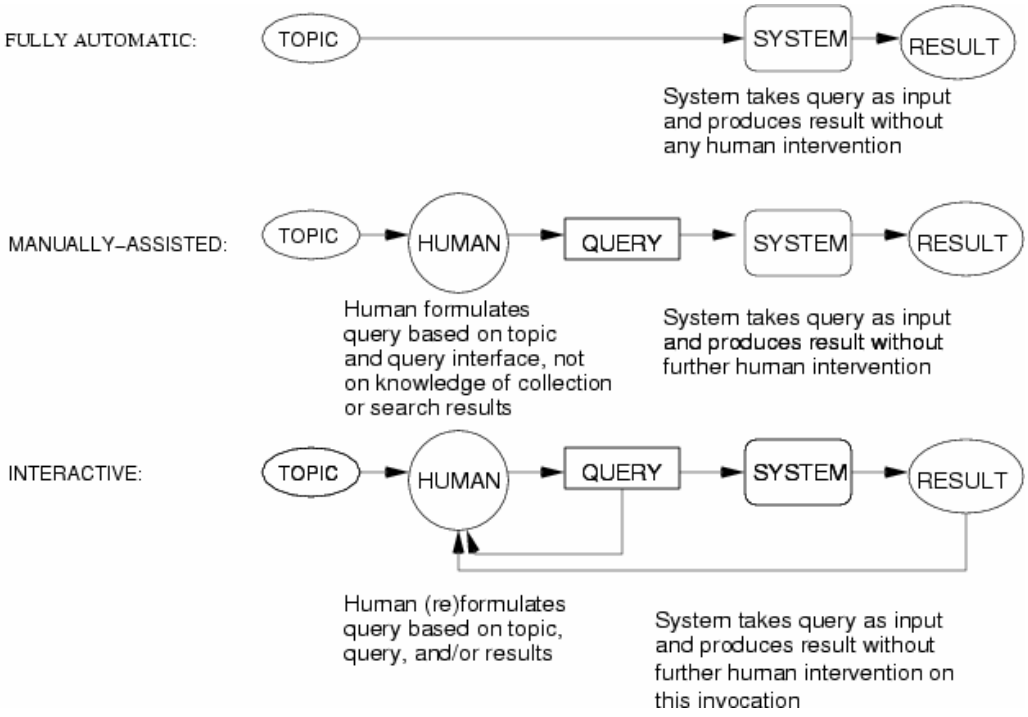


Figure 7.1 TRECVID search runs types

Some examples of queries in TRECVID 2005 and 2006 are: "*Find shots of Coodolezza Rice*", "*Find shots of a boat or ship*" (full set of queries is available in Appendix II). It is noted that this form of retrieval is aimed at finding shots which contain the required visual aspects of queries. For each issued query, a maximum of 1,000 shots is returned. The result is then evaluated by TRECVID by assessing the pool of shots gathered from all the participating systems. The assessment of shots is judged manually. For the performance measurement, TRECVID uses the mean average precision (MAP), which is suitable for evaluation in information retrieval over large corpuses where the recall rate is hard to determine.

7. 2 Performance of Video Retrieval at TRECVID

The first set of experiments is designed for testing the various components described in Chapter 6. The following subsections describe experiments created to assess be effects of: (a) query expansion and use of external parallel text corpus; (b) high level features; (c) query classification; and (d) pseudo relevance feedback.

7. 2. 1 Effects of Query Expansion and Text Baselines

The existing news video retrieval systems [Blin, Stre] use text retrieval approach for news video. We therefore establish a text only run as the baseline and a number of runs to test the following premises. First, we want to know to what extent can text-only features perform as this will provide a baseline performance for video retrieval. Second, we want to know if the window size of the text in speech transcripts (phrase level or story level) would affect retrieval performance. Third, we want to highlight the effectiveness of query expansion with the use of different types of external information sources; general web versus targeted web (parallel corpus), for query expansion. We thus carry out the following three runs by using the TRECVID 2005 and 2006 corpus:

B1) Baseline (using either phrase level or story level text)
B2) B1 with query expansion using general web
B3) B1 with query expansion using parallel news

For the phrase level runs, the complete phrases that are contained within the shot boundaries are taken as text features. Phrase information is available in the speech transcripts and machine translation outputs. For story level runs, each shot within the particular news video story will inherit all text belonging to that story. The results of the experiments are presented in Table 7.1.

Table 7.1 Retrieval performance of the text baseline in Mean Average Precision (bracket indicating improvement over respective baselines)

TRECVID2005	Using Phrase Level Text			Using Story Level Text		
	B1	**B2**	**B3**	**B1**	**B2**	**B3**
MAP	0.045	0.041 (-8.9%)	0.051 (13.3%)	0.039	0.039 (0%)	0.045 (15.4%)

TRECVID2006	Using Phrase Level Text			Using Phrase Level Text		
	B1	**B2**	**B3**	**B1**	**B2**	**B3**
MAP	0.033	0.032 (-3.0%)	0.039 (18.2%)	0.030	0.031 (6.7%)	0.033 (10%)

From Table 7.1, we can see that run B1, using only keyword matching techniques with no query expansion, could only achieve an MAP of 0.045 and 0.033 for TRECVID 2005 and 2006 respectively, indicating the poor performance of baselines.

By performing query expansion especially using the parallel news corpus, run B3 effectively improves the MAP performance to 0.051 and 0.039 based on phrase level text. This improvement is significant over the baseline and shows that the use of parallel corpus is effective. On the other hand, query expansion using general Web as in run B2 mostly yields degenerative results except for TRECVID 2006 that is based on story level text. This shows that query expansion using general Web is not effective and may even introduce noise during retrieval. As for the strange improvement in B2 in TRECVID 2006, we conjecture that it may be because the 2006 video corpus contains news from year 2005 which is closer to our "current" date. As this corpus is temporally nearer, it is possible to retrieve articles that may be more related to it during query expansion.

Another observation is that the retrieval of news video using phrase level text seems more effective than story level text. One of the main reasons we think is due to the huge amount of irrelevant shots at the story level. The retrieval based on story level text simply returns all the shots within a story without any further ranking since all shots have the same

text feature. As it is unlikely that every shot within the news video story is relevant, the precision can be greatly affected. To understand if retrieval using story level text really helps, we further check the recall performance of various runs. This is done by calculating the total number of relevant shots returned per run (using top 1000 or 2000 returned shots) as shown in Table 7.2.

Table 7.2 Recall performance: total number of relevant shots returned over 24 queries

TRECVID2005	**Using Phrase Level Text**			**Using Story Level Text**		
	B1	**B2**	**B3**	**B1**	**B2**	**B3**
# positive shots (top 1000/query)	832	830	994	899	890	1101
# positive shots (top 2000/query)	1300	1233	**1433**	1345	1300	**1501**

TRECVID2006	**Using Phrase Level Text**			**Using Story Level Text**		
	B1	**B2**	**B3**	**B1**	**B2**	**B3**
# positive shots (top 1000/query)	934	940	1099	1014	1094	1256
# positive shots (top 2000/query)	1222	1345	**1434**	1302	1543	**1599**

From Table 7.2, we further conclude that query expansion is effective since more relevant shots can be retrieved in the top 1000 and 2000 shots (per query) when comparing B3 with B1. In addition, we see that even though story level text runs yield lower MAP performance than shot retrieval runs, they have the higher total number of relevant shots retrieved. This shows that while text features are not able to pinpoint exact shots which are relevant, they can be good features for retrieving them. This is especially true since text within a story is usually coherent. The finding also suggests that there are actually plenty of positive shots found within the top 2000 shots and the performance can be greatly improved if they can be ranked correctly.

7. 2. 2 Effects of Query High Level Features

We introduce a series of techniques in Chapter 6 to incorporate HLF into the retrieval. To access the effects brought about by the use of lexical dictionary WordNet and external resource Flickr, we carry out the following re-ranking runs on top of B3, the best performing text runs. The runs are designed in an incrementally manner to access the increase in performance with respect to the addition of various components.

H1) B3 with simple query-HLF matching
H2) B3 with WordNet
H3) B3 with Flickr
H4) B3 with WordNet and Flickr

H1 is taken as the baseline run which uses keyword matching on query terms to the HLF. The HLF will be used for re-ranking if and only if the query contains terms which overlap with the HLF. For example the query "*Find shots of car on the road*" will trigger the use of HLF "*car*".

Table 7.3 Retrieval performance using HLF (bracket indicating improvement over respective H1 run)

TRECVID2005	**B3(Phrase level text, initial MAP=0.051)**				**B3(Story level text, initial MAP=0.045)**			
	H1	**H2**	**H3**	**H4**	**H1**	**H2**	**H3**	**H4**
MAP	0.081	0.084 (3.7%)	0.087 (7.4%)	0.089 (9.9%)	0.080	0.084 (5%)	0.088 (10%)	0.091 (14%)

TRECVID2006	**B3(Phrase level text, initial MAP=0.039)**				**B3(Story level text, initial MAP=0.033)**			
	H1	**H2**	**H3**	**H4**	**H1**	**H2**	**H3**	**H4**
MAP	0.052	0.053 (1.9%)	0.056 (7.7%)	0.059 (12%)	0.052	0.055 (5.8%)	0.055 (5.8%)	0.060 (15%)

From Table 7.3, we see that when HLFs are integrated into the text-only system, the improvement in MAP is significant (comparing B3 to H4: from 0.051 to 0.089 and 0.045 to 0.091 for TRECVID 2005; and from 0.039 to 0.058 and 0.033 to 0.06 for TRECVID 2006).

This validates earlier work that HLFs play a significant role in shot retrieval. The primary reason is because textual features alone are not reliable enough to pinpoint shots that are relevant to the query. The jump in performance with respect to the best text run is almost double across both datasets.

Both WordNet and Flickr are effective as improved performance can be seen in runs H2 to H4 as compared to their counterpart run H1. In addition, the run that fuses both WordNet and Flickr, run H4, yields the best performance. This indicates that both components can complement one another.

One other important observation from Table 7.3 is that the best run for both datasets comes from the run H4 that is based on story level text. This finding shows that the use of story level text is important.

Zooming into the individual performance of queries, we see that queries that can directly make use of the 50 available HLFs improve the most. Such queries are "*Find shots containing cars*", "*Find shots containing scene of fire or explosion*". This accounts for the biggest jump in improvement, from run H1 to run B3. One hypothesis that can be drawn immediately is: if there are more available HLFs, performance can be even better. On the other hand, queries which can indirectly make use of the high level features such as "*Find shots of helicopter*" benefit from the use of WordNet inference and Flickr statistics since they are able to relate high level feature like "*sky*" to help during retrieval.

The current series of runs uses the combined HLF detection results from Rank Fusion in Chapter 4. To investigate further into how detection accuracies of HLFs detectors can affect retrieval performance, we further designed another three runs having difference in HLF detection accuracies:

HS1) Using only results directly from the single best detector

HS2) Using Score Fusion from Chapter 4

HS3) Using Rank Fusion from Chapter 4

Similarly, we carry out the following re-ranking runs on top of B3 with the use of both Wordnet and Flickr. The main difference between these HS runs and H runs is that this set of runs uses only the 39 HLFs in the LSCOM-lite.

Table 7.4 HLF detection accuracies and retrieval performance (bracket indicating improvement over HS1 run)

TRECVID2005	**HS1:** Best detector	**HS2:** Score fusion	**HS3**: Rank Fusion
HLF MAP	0.351	0.398 (+13%)	0.443(+26%)
Retrieval MAP	0.083	0.087(+4.8%)	0.090(+8.4%)

TRECVID2006	**HS1:** Best detector	**HS2:** Score fusion	**HS3**: Rank Fusion
HLF MAP	0.261	0.308(+18%)	0.333(+28%)
Retrieval MAP	0.051	0.055(+7.8%)	0.058(+14%)

From Table 7.4, we can see that detection accuracies have a direct correlation with the overall retrieval performance. The retrieval MAP improves as we use HLF results with better detection rates. This is reasonable as there are certain queries that use the HLFs directly and a more accurate detection means a more precise retrieval. It is therefore necessary to improve the HLF detection so as to bring about overall improvement to retrieval.

In addition, we see a slight decrease in MAP when we compare HS3 to H4 in Table 7.3 even though the specifications are similar. The only difference is that H4 uses 50 HLFs while HS3 uses 39 HLFs. This further validates our earlier argument that increasing the number of HLFs could be helpful.

7. 2. 3 Effects of Query Classification

Besides the text and HLFs, there are also various multimodal features that can play significant roles during retrieval. It is important to leverage prior knowledge on how to combine these features effectively so as to maximize retrieval performance. We create the next series of runs to establish the importance of query classification and multimodal fusion. This series of runs utilize the rest of multi-modal features such as near duplicate information, video caption and low level image features. The runs we performed include:

M1) Query class independent retrieval (single class)
M2) Query class dependent retrieval (multi-class with heuristics)
M3) Query class dependent retrieval (multi-class with GMM learning)

M1 follows the traditional way of retrieval by fusing features using a single linear fusion function. M2 uses the nine query classes proposed in Chapter 6 with heuristically determined weights following Eqn 6.15. The rationale of this run is to determine the performance of heuristically assigned weights and machined learned weights. The last run M3 uses the nine query classes with machined learned weights using GMM. All three runs use the specifications in H4 which fuse both WordNet and Flickr for HLFs ranking.

Table 7.5 Retrieval performance using query class and other multimodal features (bracket indicating improvement over respective M1 run)

TRECVID2005	**B3(Phrase level text, initial MAP=0.051), H4(MAP=0.089)**			**B3(Story level text, initial MAP=0.045), H4(MAP=0.091)**		
	M1	**M2**	**M3**	**M1**	**M2**	**M3**
MAP	0.067	0.104 (55%)	0.116 (73%)	0.066	0.111 (68%)	0.121 (83%)

TRECVID2006	**B3(Phrase level text, initial MAP=0.039), H4(MAP=0.059)**			**B3(Story level text, initial MAP=0.033), H4(0.060)**		
	M1	**M2**	**M3**	**M1**	**M2**	**M3**
MAP	0.044	0.067 (52%)	0.070 (59%)	0.045	0.070 (56%)	0.072 (60%)

Table 7.5 tabulates the results of the runs. The best runs (MAP of 0.121 and 0.072 for TRECVID 2005 and 2006 respectively) come from M3 which uses the fusion parameters obtained through machine learning. This performance when compared with their counterparts runs B3 and H4 conform that multimodal features are indeed useful.

The results from the Table also show that query classification is also very important. As compared to M1 runs using single query class, M2 and M3 runs that use multiple classes yield significantly better results. This is mainly due to the fact that different queries require different features for evidence. For example: PERSON queries usually need more emphasis on speech and video captions while SPORT queries rely more on image level features. One interesting phenomenon we observe is that the performance of M1 runs is lower than H4 which only uses HLF. This shows that the use of more features does not necessarily mean better performance if they are not fused properly.

Table 7.6 Performance of MAP at individual query class level (using run H4 and M3 based on story level text only)

2005 Class	PERSON	SPORTS	FINANCE	WEATHER	DISASTER	HEALTH	POLITICAL	MILITARY	GENERAL
H4	0.323	0.183	NA	0.101	0.133	NA	0.101	0.012	0.023
M3	0.392	0.225	NA	0.113	0.139	NA	0.105	0.014	0.022

2006 Class	PERSON	SPORTS	FINANCE	WEATHER	DISASTER	HEALTH	POLITICAL	MILITARY	GENERAL
H4	0.263	0.146	NA	0.032	0.088	0.070	0.044	0.009	0.019
M3	0.324	0.193	NA	0.037	0.091	0.069	0.046	0.01	0.020

By analyzing the improvement at the individual query class level as in Table 7.6, we find that PERSON and SPORTS queries perform much better than the other classes. This can be attributed to two main reasons. First, these queries rely most heavily on text features. It is evident that even though speech cannot fully determine visual concepts, it does have

strong discriminating effects. Second, we find that there are certain available video features which can augment the retrieval of these particular classes of queries nicely. For example: video captions are very useful for locating a news subject which therefore provides good support for PERSON related queries. The image level features are very suitable for SPORTS queries since soccer field, tennis courts or even basketball courts are color coded. In contrast, the other queries from other classes do not benefit much from other features and mostly rely on HLFs.

7. 2. 4 Effects of Pseudo Relevance Feedback

This sub-section introduces another set of runs which uses the top *n* (n=15) shots from the result of previous M runs for PRF:

PM1) Query class independent retrieval (single class) with PRF

PM2) Query class dependent retrieval (multi-class with heuristics) with PRF

PM3) Query class dependent retrieval (multi-class with GMM learning) with PRF

Table 7.7 Retrieval performance before and after pseudo relevance feedback

TRECVID2005	**B3(Phrase level text, initial MAP=0.051)**			**B3(Story level text, initial MAP=0.045)**		
Original MAP	M1 0.067	M2 0.104	M3 0.116	M1 0.066	M2 0.111	M3 0.121
	PM1	**PM2**	**PM3**	**PM1**	**PM2**	**PM3**
MAP after Feedback	0.065	0.110	0.121	0.064	0.119	0.126
% improved	**-3.0%**	**5.8%**	**4.3%**	**-3.0%**	**7.8%**	**4.0%**

TRECVID2006	**B3(Phrase level text, initial MAP=0.039)**			**B3(Story level text, initial MAP=0.033)**		
Original MAP	M1 0.044	M2 0.067	M3 0.070	M1 0.044	M2 0.07	M3 0.072
	PM1	**PM2**	**PM3**	**PM1**	**PM2**	**PM3**
MAP after Feedback	0.044	0.068	0.072	0.043	0.073	0.076
% improved	**0%**	**1.5%**	**2.8%**	**-2.3%**	**4.3%**	**5.6%**

From Table 7.7, we observe that the use of PRF can improve or degrade the MAP

performance. The best run in this set of experiment comes from PM3 which yield a MAP of 0.126 and 0.076 for TRECVID 2005 and 2006 respectively. This performance is equivalent to the best reported performing result for TRECVID 2005 and the second best reported performing result for TRECVID 2006. This performance illustrates the system's capability and robustness.

The results found that PM1 runs obtain degenerative performance or no change with respect to M1. This can be explained as M1 runs generally have lower precisions and that means there are more negative shots near the top of the rank list. On the other hand, we observe slight improvement in PM2 and PM3 runs over their respective predecessor.

By analyzing the performance against query classes, we notice that PERSON, SPORT and DISASTER queries tend to benefit the most from PRF while other classes have mixed performance. In particular, GENERAL-class queries tend to obtain slightly worse results, partly due to poor quality of the initial retrieval.

At the individual query level, we observe that queries with good initial retrieval tend to have only slight improvement, while queries which are bad usually have no effects (since their precision is already near zero). The bulk of improvement comes mostly from middle-range performing queries. We find that 17 out of 24 queries experience improvement from TRECVID 2005 and 16 out of 24 queries for TRECVID 2006. One way to further improve the overall improvement is to perform PRF only on the set of better performing queries. In this manner, we can minimize the errors propagated through feedback. However, the research issue here is how to determine this set and the conditions in which PRF will perform well.

7. 3 Performance of Event-based Topic Browsing

The second set of experiments is the user-based survey aiming at gathering the users' responses after they have tested the proposed hierarchical browsing and topic evolution browsing. We gather a group of 15 students who have experience with online news or news video retrieval. As both news video corpuses are dated some time ago, the students are first given a brief summary of the news events which happened in the respective period. This summary is taken from Wikipedia news [Wiki] which provides a daily report of key news. Some examples of key news which happens in the targeted period is the US Presidential Election, conflicts in Iraq, Ukraine Presidential Election, North Korea Nuclear issues, news on terrorism (the full list is provided in Appendix III). The students will formulate the query themselves and carry out retrieval with the two different browsing techniques. Each student is required to enter at least 10 to 15 topics of their choice, after which they will answer the following eight questions on a scale from 1-5 (1–Strongly Disagree, 2–Disagree, 3–neutral, 4–Agree, 5–Strongly Agree).

Table 7.8 Summary of survey gathered on 15 students

User ratings	**1**	**2**	**3**	**4**	**5**
1) Quality of the retrieved results	0	0	2	6	7
2) The topic evolution browsing is easy to apprehend and use	1	0	3	5	6
3) The topic evolution browsing is flexible	0	0	0	5	10
4) The display of results in topic evolution manner helps in locating interesting articles which I may want	0	0	0	2	13
5) The topic evolution display is more effective than tradition listing	0	0	0	6	9
6) The hierarchical browsing can facilitate overall searching	0	2	5	6	2
7) The clusters contain sensible results	1	1	1	7	5
8) The use of hierarchical browsing is better than traditional listing	0	0	1	9	5
Total	2	3	12	46	57

The survey results from the 15 students are tabulated in Table 7.8. Question 1

provides a gauge on how the students generally feel about the quality of retrieval. Questions 2 to 5 are directed at the topic evolution browsing and questions 6 to 8 are directed at the use of the hierarchical browsing. The questions are designed in such a way that a higher score will indicate positive preference toward the retrieval. From the Table, we can see that most of the user scores are indicative of positive preference towards the use of our retrieval system and browsing methods. All users except one have no problems on using the topic evolution browsing display for searching news. From the collective results in Question 3, users also like the idea and flexibility of allowing user to change the time period of viewing and level relevancy against interestingness. In addition, Question 4 which is concerned with "interestingness" obtained the best response. This shows that the system is able to present events of interests to the users.

As for the questions regarding hierarchical browsing, the users also recognize that getting related articles has become easier as they are able to get them within clicks. However, some users feel that the quality of the clustering can be improved as we see from Question 7. Lastly, from Questions 5 and 8, we conjecture that almost all 15 users think that the two new browsing techniques can complement the traditional listing way of browsing by providing more valuable information in which retrieval can be done in a more efficient way.

7. 4 Performance of Event-based Video Question Answering

The last set of experiments aims to test the system's performance in particular to question answering with the answer selection technique and use of the query topic graph. We employ a total of 150 questions related to the chosen news video topics which consists of 102 context-oriented queries and 48 visual-oriented queries (evenly distributed between

TRECVID 2005 and 2006). The 102 context-oriented queries were questions modified from past TREC [Trec] Question Answering task, while the 48 visual oriented queries were queries used in the search task of TRECVID 2005 and 2006. A partial list of the 102 queries is given in Figure 7.2 (with the full list in Appendix IV).

1) <Topic: Serial Killer>
   What is the name of the serial killer?
   When was the serial killer captured?
2) <Topic: Olympics>
   Which countries are competing for Olympic 2012?
   Which city is hosting Olympic 2008?
3) <Topic: Osama bin Laden>
   Where is Osama bin Laden hiding?
   How much money is Osama bin Laden worth?
4) <Topic: Stanely Cup>
   Which team won the Stanely Cup?
5) <Topic: NBA>
   Which team won the NBA title for year 2005?
   Who is voted the best player in NBA for year 2005?
   What is the result of the match between NBA Chicago Bulls and 76ers?
6) <Topic: Hong Kong>
   What is Hong Kong unemployment rate?
   How much is the GDP of Hong Kong?
7) <Topic: US>
   What is the US consumer price index?
8) <Topic: AIDS>
   What is the name of the new drug that fights AIDS?

Figure 7.2 Partial list of questions, (1-4 for TRECVID 2005, 5-8 for TRECVID 2006)

7. 4. 1 Context-oriented Question Answering

The assessment of context questions is based on verifying if the returned video segment (in a window of 15 seconds or 30 seconds interval) contains the answers. To understand the effects brought by the answer selection technique proposed in Chapter 6, the following two runs are conducted.

C1) Question answering (baseline, base on news video story retrieval)
C2) C1 with answer selection (Eqn 5.4)

The first run C1 employ only text retrieval (Eqn 5.4) based on the query and the news video stories. The news video story with the highest score is chosen as the answer. The second run C2 uses Eqn 6.19 to determine the highest scoring answer candidates in the set of initial retrieved documents by C1. The results are tabulated in Table 7.9.

Table 7.9 Performance of context-oriented question answering (51 queries each corpus)

TRECVID 2005	Run	# Queries with correct answers	Prec.	TRECVID 2006	Run	# Queries with correct answers	Prec.
15 Seconds	**C1**	15	0.29	15 Seconds	**C1**	12	0.24
	C2	33	0.65		**C2**	30	0.57
30 Seconds	**C1**	21	0.41	30 Seconds	**C1**	14	0.28
	C2	39	0.77		**C2**	32	0.63

Combine (2005 & 06)	Run	# Queries with correct answers	Precision
15 Seconds	**C1**	27	0.265
	C2	62	0.608
30 Seconds	**C1**	35	0.343
	C2	71	0.696

From the Table, it is evident that answer selection is important as a significant improvement can be seen from C1 to C2. The system is able to obtain a best combined precision of 0.696. We found that the performance at the 30 seconds interval is better that at the 15 seconds interval. This is due to the correct answers appearing after the wrong ones especially for video segments which contain a large number of entities of the correct type.

In addition, the question answering performance seemed to be better on TRECVID 2005 than on TRECVID 2006 corpus. This can be explained as the TRECVID 2006 corpus contains a higher percentage of non-English videos and thus affects the quality of text features which is essential for the task.

7. 4. 2 Context-oriented Topic-based Question Answering

To investigate the effects from the use of the topic query graph, we conduct

additional run: C3 which uses the induced topic from the query (from Section 6.6.2) and the set of documents found through the query topic graph. The answer candidates from this run are then ranked using Eqn 6.19. The results are tabulated in Table 7.10 with C2 from previous Section for comparison.

C3) Retrieval using induced topic and query topic graph with answer selection (Eqn 6.19)

Table 7.10 Performance of context-oriented question answering with use of a query topic graph (51 queries each corpus)

TRECVID 2005	Run	# Queries with correct answers	Prec.	TRECVID 2006	Run	# Queries with correct answers	Prec.
15 Seconds	**C2**	33	0.65	15 Seconds	**C2**	29	0.57
	C3	37	0.73		**C3**	30	0.59
30 Seconds	**C2**	39	0.77	30 Seconds	**C2**	32	0.63
	C3	42	0.82		**C3**	35	0.69

Combine (2005 & 06)	Run	# Queries with correct answers	Precision
15 Seconds	**C2**	62	0.608
	C3	67	0.657
30 Seconds	**C2**	71	0.696
	C3	77	0.755

From Table 7.10, we observe that C3 runs outperform C2 runs. In comparison, C3 yields a combined precision of 0.755 while C2 only yields a precision of 0.696 at the 30 seconds segment level.

By analyzing questions in C2 and C3 which have correct answers, we found that C3 is superior due to the better quality of the initial set of retrieved news stories. This shows that the use of query topic graph is effective in gathering relevant documents and validates our earlier arguments on the use of the event model.

7. 4. 3 Visual-oriented Topic-based Questions Answering

For assessing the performance of visual-oriented questions with query topic graph on

a comparative scale, we follow the evaluation standards as in TRECVID automated search task. Three runs are designed as follows:

V1) Best run from Section 6.4 as baseline (run PM3)

V2) V1 with query topic graph (see Section 6.5)

V3) V1 with expanded topic graph (see Section 6.7)

Table 7.11 Question answering performance using a query topic graph (bracket indicating improvement over respective V1 run)

TRECVID 2005	V1	V2	V3
MAP	0.126	0.128(1.6%)	0.133(5.6%)

TRECVID 2006	V1	V2	V3
MAP	0.076	0.078(2.6%)	0.081(6.6%)

Combine 2005 & 2006	V1	V2	V3
MAP	0.101	0.103(2.0%)	0.107(5.9%)

From Table 7.11, we see that the V2 and V3 runs which use a query topic graph obtain better performance. The best combined MAP performance across both corpuses is 0.107. Run V3 that uses the expanded query topic graph, allows more relevant shots to be discovered and yield an MAP performance of 0.133 and 0.081 respectively for TRECVID 2005 and 2006. This observation confirms our earlier hypothesis that event clustering and threading can help improve retrieval. From analyzing respective queries, we see that the queries which are related to people, events and locations tend to have better responses than those targeted at objects.

Chapter 8

Conclusions and Future Work

This chapter provides the summary of the major research results presented in this thesis and discusses future directions for news video retrieval.

8. 1 Summary

In recent years, we see that the research community of multimedia retrieval is gradually shifting from analyzing one media source at a time to exploring the combination of diverse knowledge sources. The new challenge faced by the multimedia community is how to acquire and combine such diverse multimedia knowledge sources. While considerable effort has been expended on extracting valuable semantics from targeted multimedia data, less attention has been given to the problem of utilizing external resources around such data and finding a suitable strategy to fuse them. In this work, we make extensive use of multimedia technology to search across massive amount of multimedia data with the help of various external resources to find documents relevant to an information need. The contribution of this thesis can be summarized as follows.

First, this thesis described how external knowledge can be used in supporting various parts of the event-based retrieval model. In particular, the four proposed resources are language resource, image repository resource, parallel news resource and news blog

resources. We discussed several novel approaches like temporal hierarchically clustering of multi-source news articles and video information based on event entities; blog analysis for key event detection; and combining language resource and image repository for inference of query HLF in a query dependent manner.

Second, this thesis presented a news video retrieval framework that combines diverse knowledge sources using an event-based model. This model intelligently incorporated multiple sources of information from the original video as well as various external resources. Experimental results performed on the search task of TRECVID 2006 and 2007 datasets showed that the system can achieve performance better than the best reported results.

The thesis also developed user applications such as event-based topic retrieval and question answering. Besides listing the search result in a top-down manner, the system enhances news video browsing using the discovered event topic hierarchy which can effectively browse through key events and support precise question answering. Encouraging feedbacks from users show that event-based topic browsing is effective and appealing.

8. 2 Future Work

The proposed combination of external knowledge opens new opportunities to enhance retrieval given the limited amount of extractable semantics from news video. However, the approaches we described in this thesis only reveal a small tip of the full potentials of utilizing such widely available resources. Many interesting future research directions can be explored on how to integrate other types of resource to support a more precise and personalized retrieval. Future works in the pipeline includes getting users to

actively participate in the retrieval; or even providing summaries across multiple news sources. In particular, we describe two interesting works which can be built upon the retrieval engine discussed in the thesis.

8. 2. 1 Moving towards interactive retrieval

The thesis mainly discussed on techniques of fully automated retrieval. As pointed by multimedia experts, retrieval can also be viewed as a two way process since users can play the role of "trainer" in the retrieval system. It is therefore important to incorporate interactivity to promote better communication and interactivity between the user and the system.

Figure 8.1 Interactive news video retrieval user interface

One interesting direction will be looking at how to effectively fuse the automated search framework with interactive feedback. We plan to design intuitive user interfaces which can allow users to provide feedback seamlessly like in Figure 8.1. The future research will

include: (a) finding intuitive input methods such as fast keystroke actions; (b) discovering effective ways to present information to the user (providing quick previews, keyframes, etc); and (c) incorporating efficient active learning.

8. 2. 2 Personalizing summaries for story retrieval

A sample summary extracts appropriate visual segments and audio so that it can be both informative and interesting. This suggests that we should pack as much variety of news video information within the news story as possible as shown in Figure 8.2. In addition, information coherence must also be ensured during combination.

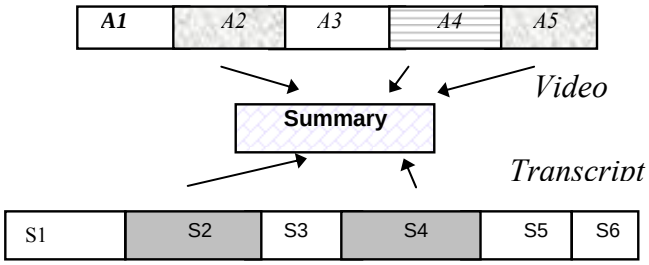


Figure 8.2 News video summarization

Given the massive amount of news information from multiple sources, it might be more appropriate for general users to obtain a summary instead of viewing every news video for obtaining information. Four major research points for generating a personalized news video summary could be how to: (a) summarize a long news video story effectively using its visual and audio content; (b) summarize across multiple news video stories; (c) understand what aspects different users might be interested in; and (d) maintain coherence of the returned segments at the same time.

References

[Adco05] J. Adcock, M. Cooper, A. Girgensohn, and L. Wilox. “Interactive Video Search Using Multilevel Indexing” CIVR 2005, Singapore, 205-214, July 2005.

[Alla98] J. Allan, R. Papka and V. Lavrenko, “On-Line New Event Detection and Tracking” SIGIR 1998, Melbourne, Australia, 37-45, 1998.

[Amir03] A. Amir, W. Hsu, G. Iyengar, C.-Y.Lin, M. Naphade, A. Natsev, C. Neti, H. J. Nock, J. R. Smith, B. L. Tseng, Y. Wu, and D. Zhang. IBM research TRECVID- 2003 video retrieval system. In NIST TRECVID-2003, Nov 2003.

[Amir04] A. Amir, J.O. Arillander, M. Berg, S.F. Chang, W. Hsu, G. Iyendar, J. R. Kender, C.Y Lin, M. Naphade, A. Natsev, J. R. Smith, J. Tesic, G. Wu, R. Yan and D. Zhang, IBM Research TRECVID-2004 Video Retrieval System. In the Notebook Paper, 82-91, TRECVID 2004.

[Amir05] A. Amir, G. Iyengar, J. Argillander, M. Campbell, A. Haubold, S. Ebadollahi, F. Kang, M.R. Naphade, A.P. Natsev, J.R Smith, J. Tesic, T. Volkmer, “IBM research TRECVID- 2005 video retrieval system” TRECVID 2005 Workshop, NIST, USA Nov 2005.

[Amit01] Singhal, Amit (2001). "Modern Information Retrieval: A Brief Overview". Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): 35-43.

[Blin] Blinkx, http://www.blinkx.com

[Bohm04] Böhm C., Kailing K., Kriegel H.-P., Kröger P.: Density Connected Clustering with Local Subspace Preferences, Proc. 4th IEEE Int. Conf. on Data Mining (ICDM'04), Brighton, UK, 2004, pp. 27-34.

[Bril01] Brill, Eric, J. Lin, M. Banko, S. T. Dumais, and Y. Ng. 2001. Data-intensive question answering. In TREC.

[Brin98] S. Brin,L. Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine. WWW7 / Computer Networks 30(1-7): 107-117 (1998)

[Bush45] Vannevar Bush, "As We May Think" - A Celebration of Vannevar Bush's 1945 Vision, at Brown University

[Cach07] http://www.cachelogic.com

[Cohe98] W. W. Cohen, R. E. Schapire, and Y. Singer. Learning to order things. In NIPS ’97: Proceedings of the 1997 conference on Advances in neural information processing systems, pages 451–457, Cambridge, MA, USA, 1998. MIT Press.

[Chai02] L. Chaisorn, T.S. Chua and C.H. Lee, “The segmentation of news video into story units” ICME 2002, Ischia, Italy, Jul 2002.

[Chai03] L. Chaisorn, C.-K. Koh, Y.-L. Zhao, H.-X. Xu, T.-S. Chua, T. Qi. Two-Level Multi-Modal Framework for News Story Segmentation of Large Video Corpus. In TRECVID 2003 Workshop, 129-134, Nov 2003.

[Chen04] M. Y. Chen and A. Hauptmann. “Searching for a specific person in broadcast news video”. Proc. of the Int'l Conf on Acoustic, Speech and Signal Processing, Vol. 3, 1036-1039. May 2004.

[Chri02] M.G. Christel, A.G. Hauptmann, H.D. Wactlar and T.D. Ng, “Collages as Dynamic Summaries for News Video” ACM Multimedia 2002, 561-569, Juan-les-Pins, France, Dec 2002.

[Chu04] Chu-Carroll, Jennifer, K. Czuba, J. Prager, A. Ittycheriah, and S. B. Goldensohn. 2004. IBM’s PIQUANT II in TREC 2004. In TREC.

[Chua04] T.-S. Chua, S.-Y. Neo, K.-Y. Li, G. Wang, R. Shi, M. Zhao and H. X. Xu, TRECVID 2004 Search and Feature Extraction Task by NUS PRIS. In TRECVID 2004, NIST, Gaithersburg, Maryland, USA, 15-16 NOV 2004.

[Chua05] T.-S. Chua, S.-Y. Neo, H.-K. Goh, M. Zhao, Y. Xiao, G. Wang, “TRECVID 2005 by NUS PRIS” TRECVID 2005 Workshop, NIST, USA Nov 2005.

[Cui05] H. Cui, M.-Y. Kan and T.-S. Chua, Generic Soft Pattern Models for Definitional Question Answering, In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development of Information Retrieval (SIGIR 2005), Salvador, Brazil, August 15 -19, 2005.

[Dasg99] S. Dasgupta (1999). "Learning Mixtures of Gaussians". Proc. of Symposium on Foundations of Computer Science (FOCS).

[Demp77] A. Dempster, N. Larid, and D. Rubin. Maximum Likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society series, 39(1):1-38, 1977.

[Echi03] Echihabi, Abdessamad and D. Marcu. 2003. A noisy-channel approach to question answering. In ACL ’03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, pages 16–23, Morristown, NJ, USA.

[Este97] Ester M., Kriegel H.-P., Sander J., Xu X.: Density-Connected Sets and their Application for Trend Detection in Spatial Databases, Proc. 3rd Int. Conf. on Knowledge Discovery and Data Mining (KDD'97), Newport Beach, CA, 1997, pp. 10-15.

[Fell98] C. Fellbaum. WordNet: An Electronic Lexical Database.MIT Press 98.
[Flickr] Flickr, http://www.flickr.com

[Fole05] C. Foley, C. Gurrin, G. Jones, H. Lee, S. McGivney, N.E. O'Connor, S. Sav, A.F. Smeaton, P. Wilkins, "TRECVid 2005 experiments at dublin city university," TRECVID 2005 Workshop, NIST, USA Nov 2005.

[Freu97] Y. Freund and R. E. Schapire, "A Decision-theoretic generalization of online-learning and an application to boosting". Journal of Computer and System Sciences, Vol. 55, no. 1, 119-139, August 1997.

[Gaug03] G. Gaughan, A. F. Smeaton, C. Gurrin, H. Lee, and K. McDonald. Design, implementation and testing of an interactive video retrieval system. In Proc. of 11th ACM MM Workshop on MIR, Nov 2003.

[Gauv02] J.L. Gauvain, L. Lamel, and G. Adda. The LIMSI Broadcast News Transcription System. Speech Communication, 37(1-2): 89-108, 2002.

[Grav02] Andrew Graves and Mounia Lalmas. Video retrieval using an mpeg-7 based inference network. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pages 339–346, 2002.

[Google] Google, http://www.google.com

[Grob05] M. Grobelnik and D. Mladenic"Visualizing very large graphs using clustering neighborhoods" Local pattern detection, Lecture notes in artificial intelligence, 3539, New York, pp. 89-97, 2005

[Hamm04] Samira Hammiche, Salima Benbernou, Mohand-Saïd Hacid, and Athena Vakali. Semantic retrieval of multimedia data. In Proc. of the 2nd ACM international workshop on Multimedia databases, pages 36–44, 2004.

[Haup96] A. Hauptmann and M. Witbrock. Informedia news on demand: Multimedia information acquisition and retrieval. In Intelligent Multimedia Information Retrieval. AAAI Press/MIT Press, Menlo Park, CA, 1996.

[Har05] S. Har-Peled, B. Sadri, "How Fast Is the k-Means Method?" Algorithmica 41, Vol3, 185-202, Jan. 2005.

[Hara00] Harabagiu, M. Sanda, I. M. Dan, P. Marius, M. Rada, S. Mihai, C. Razvan, Girju, V. R. Roxana, and M. Paul.FALCON: Boosting knowledge for answer engines. In TREC. 2000

[Haup05] A. Hauptmann., M. Christel, R. Concescu, J. Gao, Q. Jin, W.H. Lin, J.Y. Pan, S.M. Stevens, R. Yan, J. Yang, Y. Zhang, "CMU Informedia's TRECVID 2005 skirmishes" TRECVID 2005 Workshop, NIST, USA Nov 2005.
[High] Highbeam Research, http://www.highbeam.com

[Hoas04] K. Hoashi, M. Sugano, M. Naito, K. Matsumoto, F. Sugaya, and Y. Nakajima, "Shot Boundary Determination on MPEG Compressed Domain and Story Segmentation Experiments for TRECVID 2004". In the Notebook Paper, 109-120, TRECVID 2004.

[Hovy01] Hovy, Eduard, L. Gerber, U. Hermjakob, C.-Y. Lin, and D. Ravichandran. Toward semantics-based answer pinpointing. In HLT '01: Proceedings of the First International Conference on Human Language Technology Research, pages 1–7, Morristown, NJ, USA. Association for Computational Linguistics.

[Hsu05] W.H. Hsu, L. Kennedy, S.F. Chang, M. Franz, and J. Smith, "Columbia-IBM News Video Story Segmentation in TRECVID 2004", Columbia ADVENT Technical Report, New York 2005

[Hsu05b] W. H. Hsu and S.-F. Chang, "Visual Cue Cluster Construction via Information Bottleneck Principle and Kernel Density Estimation," The 4th International Conference on Image and Video Retrieval (CIVR), Singapore, July 20-22, 2005

[Hua02] X.S. Hua, P. Yin, H.J. Wang, J.F. Chen, L. Lu, M.J. Li, H.J. Zhang. MSR-Asia at TREC-11 Video Track. TREC Video Retrieval Evaluation (TRECVID 2002) 2002.

[Huur05] B. Huurnink. AutoSeek: Towards a fully automated video search system. Master's thesis, University of Amsterdam, October 2005.

[Jian00] H. Jiang, T. Lin and H.J. Zhang, "Video segmentation with the Support of Audio Segmentation and classification," ICME'2000-IEEE Int'l Conf on Multimedia and Expo, NY, USA, Jul 2000.

[Joac98] T. Joachims. Making large-scale support vector machine learning practical. In A. Smola B. Sch¨olkopf, C. Burges, editor, Advances in Kernel Methods: Support Vector Machines. MIT Press, Cambridge, MA, 1998.

[Kenn89] C. Kenneth and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography," ACL, 1989

[Kenn05] L. Kennedy, P. Natsev, and S.-F. Chang. Automatic discovery of query class dependent models for multimodal search. In ACM Multimedia, Singapore, November 2005.

[Lars99] B. Larsen and C. Aone. Fast and effective text mining using linear-time document clustering. In Proc. of the Fifth ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining, pages 16–22, 1999.

[Lee01] G. G. Lee, J. Y. Seo, S. W. Lee, H. M. Jung, B. H.Cho, C. K. Lee, B. K. Kwak, J. W. Cha, D. S. Kim,J. H. An, and H. S. Kim. 2001. SiteQ: Engineering high performance QA system using lexico-semantic pattern matching and shallow NLP. InProceedings of the 10th Text Retrieval Conference (TREC), pp. 437-446.

[Leve66] V. I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10 (1966):707–710

[Lin03] J. Lin, D. Quan, V. Sinha, K. Bakshi, D. Huynh, b. Katz, D.R. Karger, What makes a good answer? The role of context in question answering. In the Proc of the 9th International conference on human-computer interaction, pages 25-32, 2003

[Lloy82] S. P. Lloyd. Least Squares Quantization in PCM. IEEE Transactions on Information Theory, vol. 28, no. 2, pp. 129-137, 1982.

[Lowe04] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", International Journal of Computer Vision, 60, 2, pp. 91-110, 2004.

[Lscom] LSCOM Lexicon http://www.ee.columbia.edu/dvmm/lscom

[Miko05] K. Mikolajczyk and C. Schmid "A performance evaluation of local descriptors", IEEE Transactions on Pattern Analysis and Machine Intelligence, 10, 27, pp 1615--1630, 2005.

[Mold01] D. I. Moldovan and V. Rus "Logic Form Transformation of WordNet and its Applicability to Question Answering", ACL 2001, pp394-401, 2001

[Mlad98] D. Mladenic, "Machine Learning on non-homogeneous, distributed text data" PhD thesis, University of Ljubljana, Slovenia, October 1998.

[Neo05] S. Y. Neo, T. S. Chua, "Query Dependent Retrieval of News Video". In Multimedia Information Retrieval Workshop, ACM SIGIR, Brazil, Aug 2005.

[Neo06] S.Y. Neo, J. Zhao, M.Y. Kan, T.S. Chua, "Video Retrieval Using High-level features, "Exploiting Query-matching and Confidence-based Weighting", CIVR 2006, Arizona, USA, 143-152, July 2006.

[Neo06b] S-Y. Neo, Y. Zheng, T-S. Chua, Q. Tian, "News Video Search with Fuzzy Event Clustering using High-level Features" ACM Multimedia 2006, Santa Barbara, USA, 23-27 Oct 2006.

[Neo07] S.Y. Neo, Y. Ran, H.K. Goh, Y.T. Zheng , T.S. Chua, J.T. Li "The Use of Topic Evolution to help Users Browse and Find Answers in News Video Corpus", ACM MM 2007, Ausburg, Germany, Sep 2007.

[Neo07b] S.Y. Neo, Y. Zheng, H.-K. Goh, T.S. Chua, S. Tang, "News Video Retrieval Using Implicit Event Semantics," ICME 2007, Beijing, China, 2-5 Jul 2007.

[Pete05] C. Petersohn. "Fraunhofer HHI at TRECVID 2004: Shot Boundary Detection System" TRECVID 2004 Workshop, NIST, US, Nov 2005

[Poli06] R. Polikar, "Ensemble Based Systems in Decision Making" IEEE Circuits and Systems Magazine, vol.6, no.3, pp. 21-45, 2006.

[Quen04] G. M Quenot, D. Mararu, S.Ayache, M. Charhad, L. Besacier, M. Guironnet, D. Pellerin, J. Gensel and L.Carminati. CLIPS-LIS-LSR-LABRI Experiments at TRECVID 2004. In the Notebook Paper, 24-39, TRECVID 2004.

[Ravi02] Ravichandran, Deepak and E. H. Hovy. 2002. Learning surface text patterns for a question answering system. In ACL, pages 41–47.

[Raut04] M. Rautiainen and et al. TRECVID 2004 experiments at mediateam oulu. In Proc. of TRECVID, 2004.

[Resn99] P. Resnik, "Semantic similarity in a taxonomy: An information- based measure and its applications to problems of ambiguity in natural language," Journal of Artificial Intelligence Research, Nov 1999, 95–130.

[Rocc71] J. J. Rocchio. Relevance feedback in information retrieval. In The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice Hall, Englewood Cliffs, NJ, 1971.

[Ross76] Sheldon Ross. "A First Course in Probability", Macmillan, 1976.

[Rowe04] L. A. Rowe and R. Jain. ACM sigmm retreat report on future directions in multimedia research. In Proceedings of ACM Multimedia, March 2004.

[Salt75] G. Salton, A. Wong, and C. S. Yang (1975), "A Vector Space Model for Automatic Indexing," Communications of the ACM, vol. 18, nr. 11, pages 613–620.

[Sear] SearchEngineWatch, http://www.searchenginewatch.com/

[Smea03] A.F. Smeaton and P. Over. TRECVID: Benchmarking the effectiveness of information retrieval tasks on digital video. In Proc. of the Intl. Conf. on Image and VideoRetrieval, 2003.

[Smeu00] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval: the end of the early years. IEEE transactions Pattern Analysis Machine Intelligence, 22 - 12:1349 – 1380, 2000

[Smit02] J. R. Smith, C. Y. Lin, M. R. Naphade, P. Natsev, and B. Tseng. Advanced methods for multimedia signal processing. In Intl. Workshop for Digital Communications IWDC, Capri, Italy, 2002.

[Smit03] J. R. Smith. Video indexing and retrieval using MPEG-7. In B. Furht and O. Marques, editors, The Handbook of Image and Video Databases: Design and Applications. CRC Press, 2003.

[Snoe04] C.G.M. Snoek, M. Worring, J.M. Geusebroek, D.C. Koelma, and F.J. Seinstra. The mediamill trecvid 2004 semantic viedo search engine. In Proc. of TRECVID, 2004.

[Snoe05] C. G. M. Snoek, J. C. van Gemert, J.M. Geusebroek, B. Huurnink, D.C. Koelma, G.P. Nguyen, O. De Rooij, F. J. Seinstra., A.W.M. Smeulders, , C. J. Veenman., M. Worring, "The MediaMill TRECVID 2005 semantic video search engine," TRECVID 2005 Workshop, NIST, USA Nov 2005.

[Snoe05b] C.G.M. Snoek, M. Worring, and A.W.M. Smeulders. Early versus late fusion in semantic video analysis. In Proceedings of ACM Multimedia, November 2005.

[Stre] Streamsage, http://www.streamsage.com

[Techno] Technorati, http://www.technorati.com

[Tell03] Tellex, Stefanie, B. Katz, J. Lin, A. Fernandes, and G. Marton. 2003. Quantitative evaluation of passage retrieval algorithms for question answering. In SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 41–47, New York, NY, USA. ACM Press.

[Tsin03] Ch. Tsinaraki, E. Fatourou, and S. Christodoulakis. An ontology-driven framework for the management of semantic metadata describing audiovisual information. In Proc. of the 15th Intl. Conf. on Advanced Information Systems Engineering (CAiSE), 2003.

[Trec] TREC, Text Retrieval Conference, http://trec.nist.gov

[Trecvid] TRECVID, TREC Video Retrieval Evaluation, http://www-nlpir.nist.gov/projects/trecvid

[Tung06] A. Tung, R. Zhang, N. Koudas (Toronto), B. C. Ooi: Similarity Search: A Matching Based Approach. Int'l Conference on Very Large Data Bases (VLDB), Seoul, 2006

[Van02] J.-M. Van Thong, P.J. Moreno, B. Logan, B. Fidler, K. Maffey, M. Moores (2002). Speechbot: an experimental speech-based search engine for multimedia content on the web. IEEE Trans on Multimedia, Vol 4(1), 88-96.

[Vivi] Vivisimo, http://www.vivisimo.com

[Volk06] T. Volkmer and A. Natsev. Exploring automatic query refinement for text-based video retrieval. In IEEE International Conference on Multimedia and Expo (ICME), 2006.

[Voor04] E.M Voorhees. Overview of the TREC 2004 Question Answering Track. In the Notebook of the Thirteen Text Retrieval Conference (TREC 13), TRECVID 2004.
[Wiki] Wikipedia, http://www.wikipedia.com

[Wact00] H.D. Wactlar, A.G. Hauptman, M.G. Christel, R.A. Houghton, and A.M. Olligschlaeger, "Complementary video and audio analysis from Broadcast News Archives" Comm. of ACM, Vol 43. No. 2, 42-47, Feb 2000.

[West03] T.Westerveld, T. Ianeva, L. Boldareva, A. P. de Vries, and D. Hiemstra. Combining infomation sources for video retrieval: The lowlands team at TRECVID 2003. In NIST TRECVID-2003, Nov 2003.

[West04] T. Westerveld. Using generative probabilistic models for multimedia retrieval. PhD thesis, CWI, Centre for Mathematics and Computer Science, 2004.

[Wu04] Y. Wu, E. Y. Chang, K. C.-C. Chang, and J. R. Smith. Optimal multimodal fusion for multimedia data analysis. In Proceedings of the 12th annual ACM international conference on Multimedia, pages 572–579, 2004.

[Xu96] J. Xu and W. B. Croft, ``Query expansion using local and global document analysis," in Proc. ACM SIGIR, 1996.

[Xu03] J. Xu, L. Ana, and R. M. Weischedel. 2003. TREC 2003 QA at BBN: Answering definitional questions. In TREC, pages 98–106.

[Yan03] R. Yan and A. G. Hauptmann. The combination limit in multimedia retrieval. In Proc. of the eleventh ACM international conference on Multimedia, pages 339–342, 2003.

[Yan04] R. Yan, J. Yang, and A. G. Hauptmann. "Learning Query-Class Dependent Weights for Automatic Video Retrieval". Proc. of ACM MM, New York, Oct 2004.

[Yan05] R. Yan and M. R. Naphade. Semi-supervised cross feature learning for semantic concept detection in video. In IEEE Computer Vision and Pattern Recognition(CVPR), San Diego, US, 2005.

[Yan06] R. Yan and A. G. Hauptmann. Probabilistic latent query analysis for combining multiple retrieval sources. In Proceedings of the 29th international ACM SIGIR conference, Seattle, WA, 2006.

[Yan06b] R. Yan, "Probabilistic Models for Combining Diverse Knowledge Sources in Multimedia Retrieval" PhD Thesis, 2006.

[Yang03] H. Yang, L. Chaisorn, Y. Zhao, S.-Y. Neo, and T.-S. Chua. VideoQA: question answering on news video. In Proc. of the 11th ACM MM, pages 632–641, 2003.

[Yang03b] H. Yang, T.-S. Chua, S. Wang and C.-K. Koh. Structured use of external knowledge for event-based open-domain question-answering. Proc. of SIGIR 2003, Canada, Jul 2003.

[Yang04] J. Yang, M. Y. Chen, and A. G. Hauptmann. Finding person x: Correlating names with visual appearances. In Intl. Conf. on Image and Video Retrieval (CIVR'04), Ireland, 2004.

[Ye05] S. Ye, T.-S. Chua, J. R. Kei, Clustering Web Pages about Persons and Organizations, Int'l J. of Web Intelligence and Agent Systems, vol(3), pp1-14, 2005

[Yuan05] J. Yuan, L. Xiao, D. Wang, D. Ding, Y. Zuo, Z. Tong, X. Liu, S. Xu, W. Zheng, X. Li, Z. Si, J. Li, F. Lin, and B. Zhang. Tsinghua university at TRECVID 2005. In NIST TRECVID 2005, Nov 2005.

[Zhao02]Y. Zhao, G. Karypis, "Evaluation of hierarchical clustering algorithms for document datasets" Conf Information and Knowledge Management, pp515-524,McLean, Virginia, USA, 2002

[Zhao06] M. Zhao, S.Y. Neo, H. K. Goh, T. S. Chua, "Multi-Faceted Contextual Model for Person Identification in News Video" Proc. of Multimedia Modeling (MMM) 4-6 Jan, 2006.

[Zhen06] Y.-T. Zheng, S.-Y. Neo, T.-S. Chua, Q. Tian "Fast Near-Duplicate Keyframe Detection In Large-Scale Corpus for Video Search" In IWAIT 2007, Bangkok, 8-9 Jan 2007.

Publications by Main Author arising from this Research

Main Authored:

Shi-Yong Neo, Yuanyuan Ran, Hai-Kiat Goh, Yantao Zheng, Tat-Seng Chua, Jintao Li, "**The Use of Topic Evolution to help Users Browse and Find Answers in News Video Corpus**," ACM MM 2007, Augsburg, Germany, 23-29 Sep 2007.

Shi-Yong Neo, Yantao Zheng, Hai-Kiat Goh, Tat-Seng Chua, Sheng Tang, "**News Video Retrieval Using Implicit Event Semantics**," ICME 2007, Beijing, China, 2-5 Jul 2007.

Tat-Seng Chua, Shi-Yong Neo, Yan-Tao. Zheng, Hai-Kiat Goh, and Xiaoming. Zhang, "**TRECVID 2007 Search Tasks by NUS-ICT**", In TRECVID 2007, NIST, Gaithersburg, Maryland, USA, 06-07 Nov 2007.

Tat-Seng Chua, Shi-Yong Neo, Yantao Zheng, Hai-Kiat Goh, Yang Xiao, Sheng Tang, Ming Zhao, "**TRECVID 2006 by NUS-I^{2}R**" In TRECVID 2006, NIST, Gaithersburg, Maryland, USA, 13-14 Nov 2006.

Shi-Yong Neo, Yantao Zheng, Tat-Seng Chua, Qi Tian "**News Video Search with Fuzzy Event Clustering using High-level Features**" In ACM MM 2006, Santa Barbara, USA, 23-27 October 2006.

Shi-Yong Neo, Jin Zhao, Min-Yan Kan, Tat-Seng Chua "**Video Retrieval Using High-level features: Exploiting Query-matching and Confidence-based Weighting**" In CIVR 2006, Arizona, USA, 13-15 July 2006.

Shi-Yong Neo, Hai-Kiat Goh, Tat-Seng Chua, "**Multimodal Event-based Model for Retrieval of Multi-Lingual News Video**" In International Workshop on Advance Image Technology (IWAIT), Okinawa, Japan, 9-10 Jan, 2006.

Tat-Seng Chua, Shi-Yong Neo, Hai-Kiat Goh, Ming Zhao, Yang Xiao, Gang Wang "**TRECVID 2005 by NUS PRIS**" In TRECVID 2005, NIST, Gaithersburg, Maryland, USA, 14-15 Nov 2005.

Shi-Yong Neo, Tat-Seng Chua "**Query-dependent Retrieval on News Video**" In MMIR 2005, SIGIR 2005 workshop, Salvador, Brazil, 19 Aug 2005.

Tat-Seng Chua, Shi-Yong Neo, Ke-Ya Li, Gang Wang, Rui Shi, Ming Zhao and Huaxin Xu "**TRECVID 2004 Search and Feature Extraction Task by NUS PRIS**" In TRECVID 2004, NIST, Gaithersburg, Maryland, USA, 15-16 Nov 2004.

Shi-Yong Neo, Tat-Seng Chua, "**Searching for Multimedia News on the Web**", In the 9th Annual National Undergraduate Research Opportunities Programme Congress 2003 (NUROP '2003), NTU, SINGAPORE, 13 Sep 2003.

Co-Authored:

Yan-Tao Zheng, Shi-Yong Neo, Tat-Seng Chua, Qi Tian, "**Object-based Image Retrieval Beyond Visual Appearances**", MMM *2008*, Kyoto, Japan, Jan 2008

Huan-Bo Luan, Shi-Yong Neo, Hai-Kiat Goh, Yong-Dong Zhang, Shou-Xun Lin, Tat-Seng Chua, **"Segregated Feedback with Performance-based Adaptive Sampling for Interactive News Video Retrieval"** ACM MM 2007, Augsburg, Germany, 23-29 Sep 2007.

Huan-Bo Luan, Shi-Yong Neo, Tat-Seng Chua, Yantao Zheng, Sheng Tang, Yong-Dong Zhang, Jin-Tao Li, "**Active Learning Approach to Interactive Spatio-temporal News Video Retrieval***",* VideoOlympic Demo Workshop in conjunction with CIVR 2007, Amsterdam, Holland, 6-9 Jul 2007.

Huan-Bo Luan, Shou-Xun Lin, Sheng Tang, Shi-Yong Neo, Tat-Seng Chua "**Interactive Spatio-Temporal Visual Map Model for Web Video Retrieval**," ICME 2007, Beijing, China, 2-5 Jul 2007.

Yantao Zheng, Shi-Yong Neo, Tat-Seng Chua, Qi Tian, "**The Use of Temporal, Semantic and Visual Partitioning Model for Efficient Near-Duplicate Keyframe Detection in Large Scale News Corpus**," CIVR 2007, Amsterdam, Holland, 6-9 Jul 2007.

Yantao Zheng, Shi-Yong Neo, Tat-Seng Chua, Qi Tian "**Fast Near-duplicate Keyframe Detection in Large-scale Corpus for Video Search**" In IWAIT 2007, Bangkok, 8-9 Jan 2007.

Ming Zhao, Shi-Yong Neo, Hai-Kiat Goh, Tat-Seng Chua, "**Multi-Faceted Contextual Model for Person Identification in News Video**" In Multimedia Modeling (MMM), Beijing, China 4-6 Jan, 2006.

Hui Yang, Lekha Chaison, Yunlong Zhao, Shi-Yong Neo, Tat-Seng Chua, "**VideoQA: Question Answering on News Video**", In the Proceedings of the Eleventh Annual ACM International Conference on Multimedia (ACM MM'2003), Berkeley, California, USA, 2-8 Nov 2003.

Appendix I

List of extractable name entities

HUM_BASIC
HUM_ORG
HUM_PERSON
LOC_BASIC
LOC_ALL
LOC_CITY
LOC_COUNTRY
LOC_COUNTY
LOC_ISLAND
LOC_LAKE
LOC_MOUNTAIN
LOC_PLANET
LOC_PROVINCE
LOC_RIVER
LOC_STATE
LOC_TOWN
LOC_OCEAN
NUM_AGE
NUM_AREA
NUM_BASIC
NUM_COUNT
NUM_DEGREE
NUM_DISTANCE
NUM_DURATION
NUM_MONEY
NUM_PERCENT
NUM_RANGE
NUM_SIZE
NUM_SPEED
OBJ_ANIMAL
OBJ_BASIC
OBJ_BREED
OBJ_COLOR
OBJ_CURRENCY
OBJ_GAME
OBJ_LANGUAGE
OBJ_PLANT
OBJ_RELIGION
OBJ_WAR
TME_BASIC
TME_DAY
TME_MONTH
TME_TIME
TME_YEAR
TME_DATE

Appendix II

TRECVID 2005 Queries

0149 Find shots of Condoleeza Rice
0150 Find shots of Iyad Allawi, the former prime minister of Iraq
0151 Find shots of Omar Karami, the former prime minister of Lebannon
0152 Find shots of Hu Jintao, president of the People's Republic of China
0153 Find shots of Tony Blair
0154 Find shots of Mahmoud Abbas, also known as Abu Mazen, prime minister of the Palestinian Authority
0155 Find shots of a graphic map of Iraq, location of Bagdhad marked - not a weather map
0156 Find shots of tennis players on the court - both players visible at same time
0157 Find shots of people shaking hands
0158 Find shots of a helicopter in flight
0159 Find shots of George Bush entering or leaving a vehicle (e.g., car, van, airplane, helicopter, etc) (he and vehicle both visible at the same time)
0160 Find shots of something (e.g., vehicle, aircraft, building, etc) on fire with flames and smoke visible
0161 Find shots of people with banners or signs
0162 Find shots of one or more people entering or leaving a building
0163 Find shots of a meeting with a large table and more than two people
0164 Find shots of a ship or boat
0165 Find shots of basketball players on the court
0166 Find shots of one or more palm trees
0167 Find shots of an airplane taking off
0168 Find shots of a road with one or more cars
0169 Find shots of one or more tanks or other military vehicles
0170 Find shots of a tall building (with more than 5 floors above the ground)
0171 Find shots of a goal being made in a soccer match
0172 Find shots of an office setting, i.e., one or more desks/tables and one or more computers and one or more people

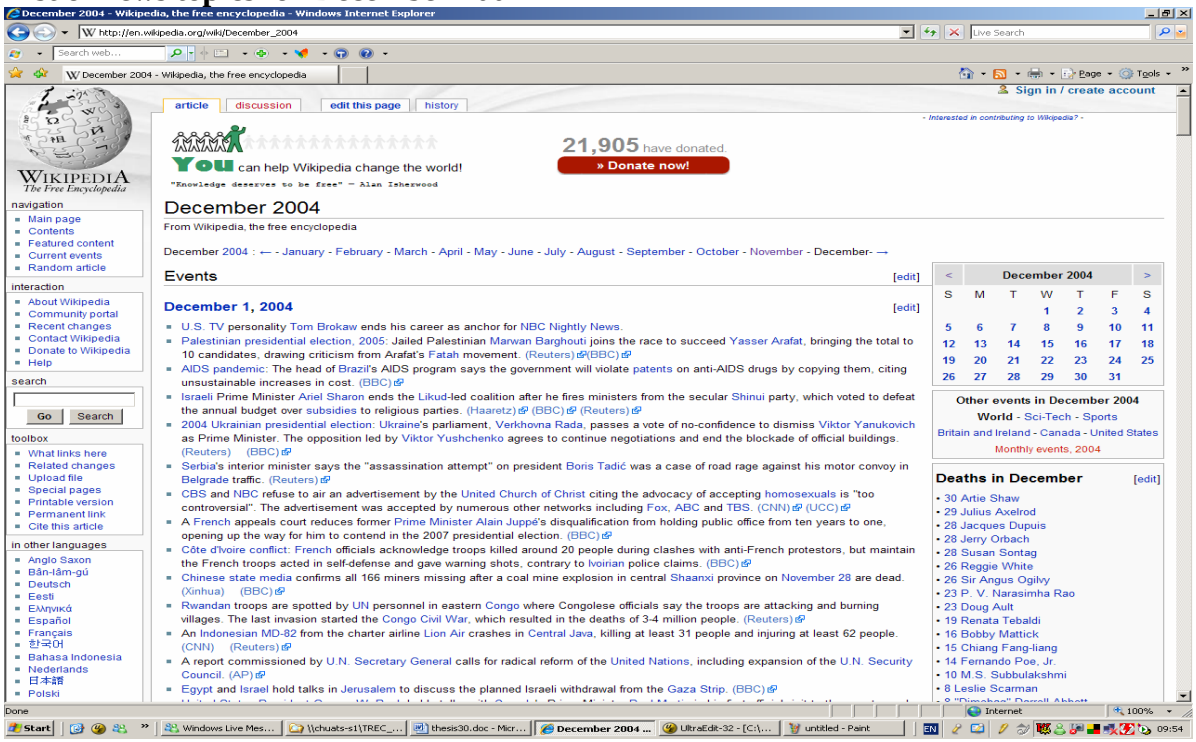
TRECVID 2006 Queries

0173 Finds shots with one or more emergency vehicles in motion (e.g., ambulance, police car, fire truck, etc.)
0174 Find shots with a view of one or more tall buildings (more than 4 stories) and the top story visible
0175 Find shots with one or more people leaving or entering a vehicle
0176 Find shots with one or more soldiers, police, or guards escorting a prisoner
0177 Find shots of a daytime demonstration or protest with at least part of one building visible
0178 Find shots of US Vice President Dick Cheney
0179 Find shots of Saddam Hussein with at least one other person's face at least partially visible
0180 Find shots of multiple people in uniform and in formation
0181 Find shots of US President George W. Bush, Jr. walking
0182 Find shots of one or more soldiers or police with one or more weapons and military vehicles
0183 Find shots of water with one or more boats or ships
0184 Find shots of one or more people seated at a computer with display visible
0185 Find shots of one or more people reading a newspaper
0186 Find shots of a natural scene - with, for example, fields, trees, sky, lake, mountain, rocks, rivers, beach, ocean, grass, sunset, waterfall, animals, or people; but no buildings, no roads, no vehicles
0187 Find shots of one or more helicopters in flight
0188 Find shots of something burning with flames visible
0189 Find shots of a group including least four people dressed in suits, seated, and with at least one flag
0190 Find shots of at least one person and at least 10 books
0191 Find shots containing at least one adult person and at least one child
0192 Find shots of a greeting by at least one kiss on the cheek
0193 Find shots of one or more smokestacks, chimneys, or cooling towers with smoke or vapor coming out
0194 Find shots of Condoleeza Rice
0195 Find shots of one or more soccer goalposts
0196 Find shots of scenes with snow

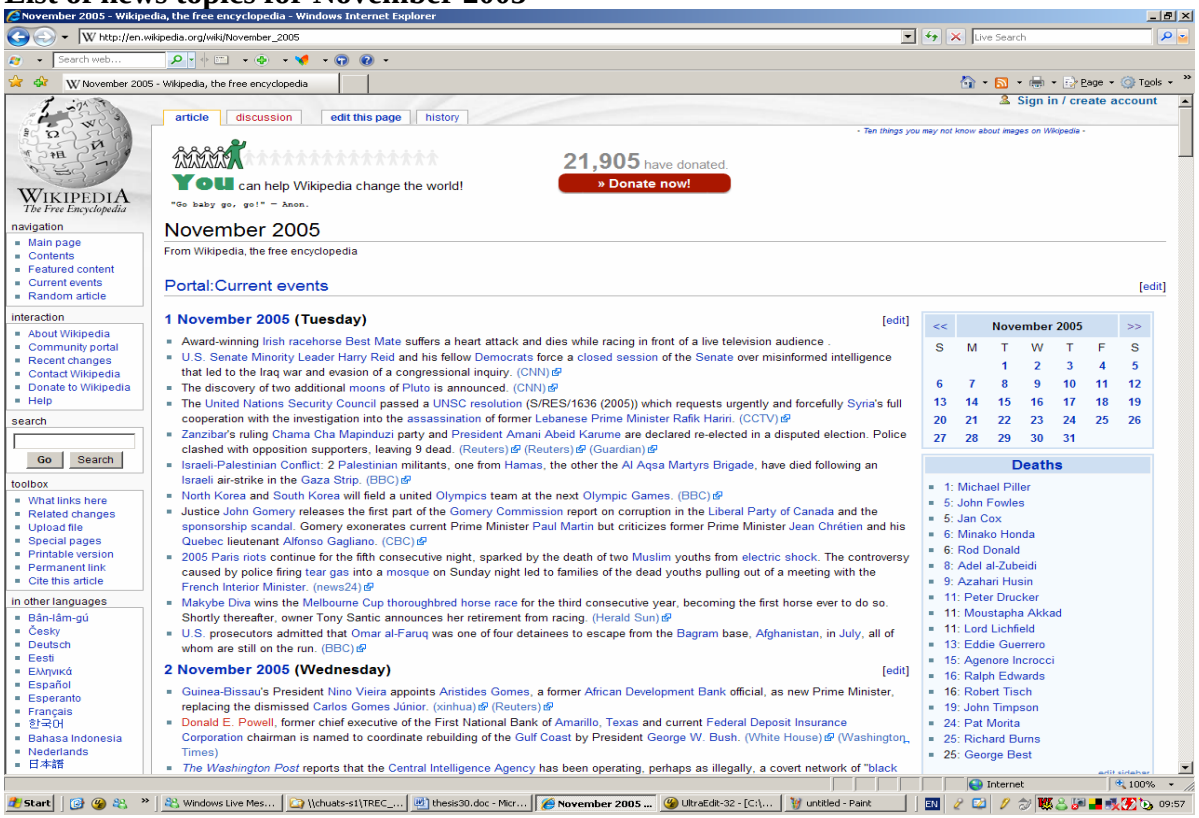
Appendix III

List of news topics for November 2004

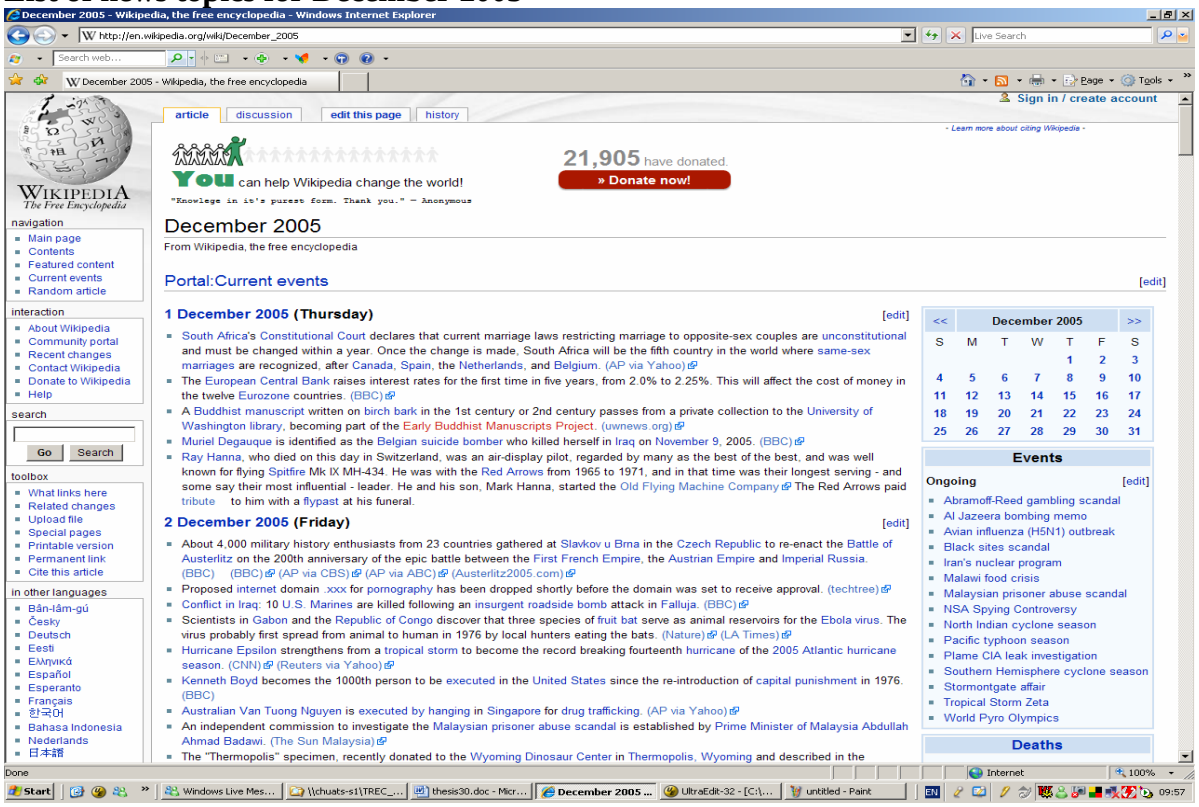
List of news topics for December 2004



List of news topics for November 2005



List of news topics for December 2005



Appendix IV

List of Queries (visual queries follows Appendix II)

What is the name of the serial killer
Which countries are competing for Olympic 2012
Which team won the Stanely Cup
Which team won the NBA title
What is Hong Kong unemployment rate
What is the US consumer price index
What is the name of the new drug that fight AIDS
What is the result of the match between Chicago Bulls and Rocketman
Which cities have nuclear power stations
When was the comet discovered
When was Wen Jiabao born
What does WTO stand for
Where do Rhodes scholars study
What kind of animal is an agouti
Who won the US presidential election
How many people are killed in the bomb attack on 3rd of November
Who is the son of Sheikh Zayed
What is their annual revenue of Microsoft
How many votes did George Bush win
Which states prohibits same sex marriages
When did Arafat die
How old is Arafat when he pass away
Which hospital did Arafat went
How much is Microsoft paying Novell
What is the magnitude of earthquake that rock Japan
Who kill Laci Peterson
How many prisoners are captured by US troops in Fallujah
How much money was raise in the 25th annual BBC Children in Need telethon
The girl who survive rabies without a vaccination is from which state
How many people die from Spanish Flu in 1920
What Las Vegas hotel was made famous by the Rat Pack
What kind of a community is a Kibbutz
In what year did the first Concorde passenger flight take place
How many seats are in the cabin of a Concorde
Where was Franz Kafka born
What nationality is Franz Kafka
What sport does Jennifer Capriati play
When did Amtrak begin operations
How many passengers does Amtrak serve annually
When was Nimitz born?
How many episode did Ken Jennings host in Jeopardy
When was the USS Constitution commissioned
Who won the Nobel Prize for Peace
Who won the Nobel Prize for Physics
Who won the Nobel Prize for Chemistry
Who is the U.S. ambassador to the Vatican
How much money is stolen from the headquarters of the Northern Bank in Belfast, Northern Ireland
What is the magnitude of earth quake which rock Macquarie Island
On what date was Bashar Assad inaugurated as the Syrian president
How many people did Yoo Young-Chul kill
How many Ecuador's Congress court justices are dismissed
What is the nationality of Azzam Azzam
Who is the culprit of the forest fire in Australia
Which country is hosting Olympic 2008

Which city is hosting Olympic 2008
What is the US consumer price index
What is the name of the company which gone listed for ten billion dollars
What are the cities affected by earthquakes
What is the name of the Irish racehorse which suffers a heart attack and dies in front of the audience
Who is the president of Guinea-Bissau
How many people die from the car bomb in India
When did the bomb exploded
Who is the president of Peru
Which country has the highest AIDS infection rate
Who is the first female president on the continent of Africa
When did Eddie Guerrero die
How many prisoners are found in the government bunker in Baghdad
When did Prime Minister of Israel, Ariel Sharon resign
When is the Battle of Austerlitz
Where is the Battle of Austerlitz carried out
What is the Democratic Republic of the Congo known as
What is the magnitude of earthquake that rocks Zaire
Where did the C-130 airplane crash
How many people were killed in the C130 crash
How much was Dreamwork SKG sold
Which company is chosen to manufacture $100 laptop
Who is the president of Brazil
How many people are detain by police at the World Trade Organization Ministerial Conference of 2005
Who is the US vice president
How many people are killed in the plane crash in Miami Beach
Who is the former president of Iraq
How many seats are taken by the Chinese Nationalist Party in the recent election