Search in Audiovisual Broadcast Archives

Bouke Huurnink

Search in Audiovisual Broadcast Archives

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof.dr. D.C. van den Boom
ten overstaan van een door het college voor
promoties ingestelde commissie, in het openbaar
te verdedigen in de Agnietenkapel
op vrijdag 26 november 2010, te 14.00 uur

door

Bouke Huurnink

geboren te Eindhoven

Promotiecommissie

Promotor:	Prof. dr. M. de Rijke
Co-promotor:	Prof. dr. ir. A.W.M. Smeulders
Overige Leden:	Dr. C. Monz
	Dr. C.G.M. Snoek
	Prof. dr. A.Th. Schreiber
	Prof. dr. A.F. Smeaton

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

SIKS Dissertation Series No. 2010-50
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

The investigations were supported by the *Mu*ltimedia a*N*alysis for *C*ultural *H*eritage (MuNCH) project, subsidized by the Netherlands Organization for Scientific Research (NWO) under the project number 640.001.501. They were additionally supported by the Center for Creation, Content and Technology (CCCT), the Dutch Ministry of Economic Affairs, and Amsterdam Topstad under the Krant van Morgen project.

Cover design by Louise Parsonage.
Production: F&N Boekservice - Amsterdam.

ISBN: 978-90-786-7599-0

Contents

Acknowledgments

Thanks go first and foremost to my advisor, Maarten de Rijke. Maarten was responsible for my first introduction to information retrieval, and has been a constant guiding light throughout the path of the PhD thesis. I have especially valued his ability to step back and take a look at the bigger picture, which has resulted in many unexpected and motivating discussions at our weekly meetings.

I would also like to thank my co-promotor, Arnold Smeulders. Though we didn't get a chance to speak often, I have come to appreciate the sometimes-unexpected nature of his viewpoints, and his insights on the nature of scientific thinking will stay with me beyond the lifespan of this thesis.

I am very grateful to Christof Monz, Guus Schreiber, Alan Smeaton, and Cees Snoek for agreeing to serve as opponents on my PhD committee.

Many people have helped contribute to the thoughts, ideas, experiments, and text in this thesis, and I am grateful to all my co-authors and collaborators. In particular I would like to thank Cees Snoek, who was there from the beginning and who takes the blame for leading me down the path to TRECVID. I would also like to thank the MediaMill team, for helping me understand how all of this multimedia analysis stuff works, and who contributed the low-level features and concept detectors for the thesis. The project members of the MuNCH team had significant input in various papers and studies, especially my two co-project employees, Laura and Michiel. I owe thanks to the non-science-type friends who have been roped into checking my writing over the years - Adam, Patty, and Shaun, I'm looking at you.

During the last months of thesis writing there was draft checking to be done, and I would like to thank Edgar, Frank, Katja, Krisztian, Ork, Xirong and Valentin for their feedback on earlier versions of this manuscript. Louise, thanks heaps for designing the awesome cover. Kwin and Merel, my trusty paranymphs, racked their brains over the Dutch version of the summary, *dank jullie wel*.

I have been lucky to be part of a vibrant and energetic research group, and I would like to thank all of the ILPSers for making the UvA a great place to work and play... and for the swimming, concerts, philosophical discussions, impromptu dinners, drinks after work, and general good times. A special mention goes to Valentin, who over the years that we shared an office together has been a source of sage advice, a patient sounding-board, and a good friend.

Beeld en Geluid was my work away from work, and I owe a big thanks to the R&D team there for making these visits an experience to look forward to. Johan van Oomen with his unflagging energy was able to move mountains to make the transaction logs that form an integral part of this thesis a reality, and Wietske van den Heuvel was a great help in teaching me the nitty-gritty of how an audiovisual broadcast archive works.

My friends, both near and far, at work and away, have been an invaluable source of support through thick and thin. One of the best motivators for finishing this thesis has been looking forward to spending more time with them.

Finally, I would like to thank my parents for their unconditional love and support.

Bouke Huurnink
Amsterdam, October 2010

Chapter 1

Introduction

As early as 1935, video archivists aimed to individually describe each shot in each film acquired by their archive. However, it soon became apparent that it was impossible to accomplish this goal through manual labor, given the limited human resources at their disposal and the large numbers of new videos coming into their archives [165]. Archivists had to settle instead for either fully describing the shots of a small number of carefully selected films, or for describing films and other types of audiovisual material at the program level, while occasionally providing more detailed within-video descriptions at the archivist's discretion. Now, the original dream of individually describing each shot in each video in an archive has come within grasp. Where *manual* description of shots is not realistic, machines may fill the gap with *automatic* shot description.

In this thesis we will walk along the intersection between the archive and the machine, studying searchers in today's audiovisual broadcast archives on the one hand, and exploring the potential for improving their searches with automatically generated descriptions of shots on the other.

The Audiovisual Broadcast Archive The professional audiovisual broadcast archive stores television and film broadcasts for use and reuse by professionals such as news producers, documentary makers, journalists and advertising agencies. Until recently broadcasts were stored on analog media, and archives consisted of large collections of film rolls, cartridges, spools and tapes in a multitude of formats. Access to the content of an audiovisual broadcast archive was limited to those physical, analog artifacts. Archive specialists, physically present at the archive, functioned both as search assistants and as gatekeepers of the archive. A customer could speak to a specialist, who would then provide the physical tapes that the customer desired.

If the customer wanted to search through videos to make a selection, he or she would have to come to the archive in person and, with the aid of the archival specialist, look for tapes and view them in a special viewing room.

In the new, digital, situation, archive customers can be provided with online access to video. The searching and purchasing process can be done with the customer neither contacting an archive specialist nor coming to the archive in person. This increases ease of access for the archive's customers. However, the burden of search is now placed squarely on the customer, who may lack the specialized knowledge of the archive contents that an archive specialist has.

Information Retrieval With the burden of search transferred from the expert archive specialist to the non-expert archive customer, information retrieval becomes an important factor in achieving a satisfactory user experience. Here multimedia broadcast archives can draw from the extensive research in the text domain. Decades before digital multimedia became widespread, digital text (which requires less storage, processing, and bandwidth than video) underwent a similar proliferation. This led to extensive research efforts into the retrieval of textual information. We see the fruits of these labors in the many commercial applications that are now available, allowing us to search hard drives, the Internet, our library catalogues, and more [24].

Audiovisual broadcast archives are applying textual information techniques to their own archival catalogs, which have in many cases been maintained for decades by professional archivists. These catalogs contain manual entries for the archive contents that consist of, at minimum, basic production data such as the broadcast name, medium, and copyright owner. There may also be extra information, for example the names of the people in the broadcast, the subject of the broadcast, different scenes of the broadcast, and sometimes even detailed descriptions of certain shots in the broadcast. By allowing customers to search on the catalog data using text retrieval techniques the archive opens up its content to them.

However, there are limitations to searching on catalog information. The exhaustiveness of the catalog entry for a particular broadcast varies according to the perceived "importance" of a broadcast, and according to the availability of annotators. Thus a broadcast with a brief catalog entry will be much less likely to be retrieved than a broadcast with a very descriptive content, even if the broadcast content is an excellent match for the searcher's information need. Automatic generation of metadata may provide an alternative to search on catalog information.

Automatically generated content metadata Recent advances in multimedia content analysis have resulted in methods for automatically describing the content

of digitized videos. An early development was the automatic detection of shot boundaries, which enables video to be broken up into coherent visual segments [49]. Another development has been *automatic speech recognition*, which allows us to automatically transcribe dialog and search it as text [11, 178, 194]. *Low-level features* such as color and texture can be computed from the video signal of a shot and used to search by example [148]. And finally *concept detectors* can automatically label video with respect to high-level semantic concepts such as *Boat*, *Bicycle*, and *Sky*; these labels can in turn be used for search [153]. While multimedia content analysis methods do not always produce results with the same accuracy as a human annotator, once they have been developed they can be applied to create annotations with no further manual labor. In addition, progress in the field is rapid, and for concept detectors at least, performance has been continually improving over the past years and is expected to keep doing so [151].

Our Setting Access to video has become ubiquitous for those with access to modern digital technology. Video material can now be accessed not only through the well-established media of broadcast television, videos, and DVDs, but also through digital services that can be accessed from PCs, laptops, web-enabled televisions, and even mobile phones. We have seen the rise of digital video for diverse purposes such as disseminating lectures, live video conferencing, and for casual entertainment and consumption. Video consumption "in the wild" is typically interspersed with regular activities in daily life [28]. In our setting we focus on video consumption, not in daily life, but rather in the setting of the audiovisual broadcast archive.

The Netherlands Institute for Sound and Vision is the national Dutch audiovisual broadcast archive, and stores over 700,000 hours of audiovisual material. When work on this thesis began, the archive had just completed a transition from storing all broadcasts in analog format to making them digitally available. As a consequence, they were interested in exploring how state-of-the-art automatically generated metadata could be applied to the new digital data and help their users search better through the archive content. This interest resulted in their participation in the Multimedia aNalysis for Cultural Heritage (MuNCH) project, which funded this thesis. The aim of the project has been to investigate how to provide faster and more complete access to videos in cultural archives through digital analysis, and the thesis has taken place in this setting.

1.1 Research Questions

In this thesis we will examine the searches of media professionals in an audiovisual broadcast archive, and then move on to investigate search with automatically

generated data, and how it might potentially improve search for professionals in an archive. Accordingly, we have two main questions that we wish to answer, the first concerning *searchers*, and the second concerning *search*.

Our first main question is, ***How do professional users search through audiovisual broadcast archives?*** To study how automatically generated content metadata may help improve retrieval effectiveness for professional users in audiovisual broadcast archives, we first need to know what kinds of retrieval tasks they perform. In other words, what do users of the archive search for? What tasks do they want to fulfill? This gives rise to our first research question,

> **RQ 1** *What kinds of content do media professionals search for, and in what manner do they search for it?*

Not only do professionals search for material in the archive, they also purchase material for reuse in new productions. We turn to explore these queries in more detail, asking

> **RQ 2** *Can we recreate those searches by media professionals that result in purchases, and use them to create an artificial testbed for retrieval evaluation?*

Having gained a deeper understanding of the professional search in audiovisual broadcast archives, we turn towards the application of automatically generated content metadata for search, and our second main question, ***How can we improve search for content in audiovisual broadcast archives?*** We start by asking two questions aimed at increasing our understanding of how retrieval with different sources of automatically generated metadata can be improved. The first concerns searching with concept detectors, where typically a retrieval system needs to select the right concept detectors to use for search:

> **RQ 3** *Given that search by visual concept detector is a valuable method for content-based video retrieval, how can we identify the correct visual concepts to use for a query?*

Our next question concerns search with automatically generated speech transcripts, specifically the temporal mismatch that occurs when using information from the audio signal to search for objects in the video signal,

> **RQ 4** *Within a video broadcast, the same object may appear multiple times within and across the video and audio signal, for example being mentioned in speech and then appearing in the visual signal. How can this phenomenon be characterized, and can we model and use this characteristic so as to improve cross-stream retrieval of visual items using transcripts?*

Finally we move from investigating search with individual types of automatically generated metadata in isolation, and move to study how they could help improve retrieval in the audiovisual archive, asking

> **RQ 5** *What is the potential impact of content-based video retrieval in the audiovisual broadcast archive, taking into account both the needs of professional users, and the manually created data already present in the archive?*

With these research questions, we aim to gain a better understanding of how search is currently done in the archive, and how (and whether) search for content may be improved using automatically generated content metadata.

1.2 Organization

This thesis is structured in two parts. Part I explores the searching and purchasing behavior of professional users of the audiovisual broadcast archive. Part II presents models and methods for improving retrieval performance in the archive. Related work and a conclusion is presented for each part.

We start Part I with background on previous efforts on understanding searchers for audiovisual material in Chapter 2. In Chapter 3 we quantify the actions of searchers at the Netherlands Institute of Sound and Vision through a large-scale transaction log analysis. This gives a characterization of the search behavior of professional users at the archive, particularly in terms of what they are searching for as well as the types of material that they buy. In Chapter 4 we examine in more detail the queries that result in purchases, using insights about how archive users formulate their queries. Here our primary goal is to gain a deeper understanding of the queries in the archive by creating a simulator that can generate queries with similar properties to the real queries in the archive. We present our conclusions on user behavior in the audiovisual archive in Chapter 5, and discuss the implications of our studies for the second part of the thesis. .

Moving on to Part II, we start by reviewing related work in content-based video retrieval in Chapter 6. We follow this by specifying the evaluation collections that we will use for our retrieval experiments in Chapter 7, including a collection designed specifically to reflect the environment and information needs of the media professionals that we studied in Part I. In Chapter 8 we examine the way in which people select concepts for video retrieval, and compare this to machine-based concept selection. In Chapter 9, we model redundancy between mentions of visual items in automatically recognized speech and their appearance in the video track and use this to improve transcript-based retrieval of visual objects in the archive. In Chapter 10 we look

towards the future and combine multiple sources of retrieval information, both automatically generated and manually created, with a view to investigating their final impact on retrieval in the archive. We present our conclusion of the implications of the work in Part II for improving search in the archive in Chapter 11.

Finally we present our overall conclusions to the thesis in Chapter 12.

1.3 Main Contributions

We group the main contributions of the thesis into two areas: understanding searchers in audiovisual archives, and improving search in audiovisual archives.

Understanding Searchers

- Our first contribution in understanding searchers is a large-scale transaction log analysis of the electronic traces left behind by media professionals at an national audiovisual broadcast archive. We provide insights into the kinds of content that these professionals are searching for, and how they search for it.
- Our second contribution in this area is a framework for simulation of searches and purchases in the audiovisual archive, as well as a method for validating such simulators. This framework and validation approach can be extended to other commercial domains where transaction logs are available.

Improving Search in the Archive

- We propose a new approach for assessing the automatic selection of visual concepts to be used in finding relevant video fragments.
- We contribute to the understanding of the temporal aspect of the occurrence of visual items in audiovisual material. We characterize the clustering behavior of said items, and develop a retrieval model that is capable of incorporating those characterizations.
- We contribute an evaluation methodology for assessing the potential impact of content-based video retrieval in the real-world context of audiovisual archives, as well as an implementation of said methodology.
- Finally, we contribute resources that were developed and used for our experiments in improving search to the research community.

1.4 Origins of the Thesis

This thesis is based in part on work published at other venues. Early versions of the work presented in Part I were published as

- "The search behavior of media professionals at an audiovisual archive: A transaction log analysis" [75] (Chapter 3),
- "Validating query simulators: An experiment using commercial searches and purchases" [74] (Chapter 4),
- "Simulating searches from transaction logs" [73] (Chapter 5).

Part II builds on work presented in

- "Today's and tomorrow's retrieval practice in the audiovisual archive" [76] (Chapters 7 and 10),
- TRECVID working notes papers [154, 157, 159, 161] (Chapter 7)
- "Term selection and query operations for video retrieval" [70] (Chapter 7),
- "Assessing concept selection for video retrieval" [72] (Chapter 8),
- "Adding semantics to detectors for video retrieval" [158] (Chapters 7 and 8),
- "Exploiting redundancy in cross-channel video retrieval" [69] (Chapter 9),
- "The value of stories for speech-based video search" [71] (Chapter 9),

In addition, the material in Part I is partly based on [61, 62] and [106]. The material in Part II is partly based on [70] and [71].

Part I
Searchers in Audiovisual Broadcast Archives

Documentary makers, journalists, news editors, and other media professionals routinely require previously recorded audiovisual material for reuse in new productions. For example, a news editor might wish to reuse footage shot by overseas services for the evening news, or a documentary maker describing the history of the Christmas tradition might desire footage from 1930s Christmas broadcasts. To complete production, the media professional must locate audiovisual material that has been previously broadcast in another context. One of the sources for reusable broadcasts is the audiovisual archive, which specializes in the preservation and management of audiovisual material [38]. Despite the fact that an increasing amount of audiovisual programming is being digitally produced, little is known about the search behavior of media professionals locating material for production purposes.

In this first part of the thesis we study the professional searchers in one of today's national audiovisual broadcast archives. Our contributions in Part I include: a description of their behavior archive in terms of sessions, queries, query terms, and purchases; and a framework for building and validating simulators of their queries and purchases that incorporate knowledge gained from the actions of real-world users.

Chapter 2

Audiovisual Searchers

In this chapter we review studies of the "human in the loop" of the video search process, and outline how they have led to and motivated the user studies that we perform in the first part of the thesis. We start with a review of some relevant aspects of user studies in the discipline of information retrieval in Section 2.1. In Section 2.2 we zoom in on studies of people searching for video material "in the wild." This is followed by a review of studies of searchers in controlled experiments in Section 2.3. Finally, we discuss the implications of the studies of audiovisual searchers for the remainder of the thesis in Section 2.4.

2.1 User Studies in Information Retrieval

Studies of information seeking go back as far as 1948, with a survey of sources used by scientists for obtaining scientific literature [138]. Following on from this, a large body of work on studies of searching behavior in general has emerged, which we do not aim to comprehensively review here. Rather, we will focus on aspects of user studies that are relevant to the thesis. Specifically we will review query typologies and transaction log analysis, which are two building blocks of the studies in Chapters 3 and 4. For more comprehensive reviews of the literature on user studies in information retrieval we refer the reader to the overviews provided by, for example, Marchionini [103] on information seeking in electronic environments and Case [17] on methodological approaches.

Query typologies An important result to come out of the increasing number of detailed studies of users' search behavior is multiple typologies of queries and searches.

Broder [10] describes three query types in the context of web search: informational ("I need to know about a topic"), navigational ("Take me to a specific item or site") and transactional ("I need to purchase or download a product or service"). This typology has served as the basis for other query classification schemes, including those by Rose and Levinson [130], Kellar et al. [90], and Jansen et al. [83]. Smeulders et al. [147] provide a categorization specific to content-based image retrieval queries: target (or known-item) search (when the user has a specific image in mind), category search (retrieving an arbitrary image representative of a specific class), and search by association (search starts with no aim other than to find interesting things). In query typologies, queries are categorized such that the searcher is expected to be satisfied by very different needs for each type. By understanding what types of queries are commonly issued in a search environment, search engines can be designed to ensure they satisfy the different types of needs.

Transaction log analysis Users can be hard to access *in situ*, and as a result the study of user interactions recorded by retrieval systems—*transaction log analysis*—has become a common method of studying retrieval behavior. Information science has a long history of transaction log analysis, from early studies of the logs created by users of library online public access catalog (OPAC) systems [120] to later studies of the logs of Web search engines [79]. This was followed by the analysis of more specialized search engines and their transaction logs. For instance, Mishne and de Rijke [108] study the use of a blog search engine through a log file analysis and Carman et al. [14] examine the difference between the vocabularies of queries, social bookmarking tags, and online documents. Three frequently used units of analysis have emerged from the body of work: the *session*, the *query*, and the *term* [79]. These three units will also play a central role in our transaction log analysis in Chapter 3.

Over the years, transaction log analysis has proved an apt method for the characterization of user behavior. Its strengths include its non-intrusive nature — the logs are collected without questioning or otherwise interacting with the user — and the large amounts of data that can be used to generalize over the cumulative actions taken by large numbers of users [78]. It is important to note that transaction log analysis faces limitations: not all aspects of the search can be monitored by this method, for example, the underlying information need [127]. Thus, we cannot say with any certainty what a user is looking for on basis of what appears in the transaction logs alone; for this, qualitative methods are required. It can also be difficult to compare across transaction log studies of different systems due to system dependencies and varying implementations of analytical methods. Comparability can be improved to some extent by providing clear descriptions of the system under investigation and the variables used [79]. When two systems are comparable, then we can study differences and similarities between the searchers using the systems.

2.2 Field Studies of Audiovisual Retrieval

In this section we review studies of audiovisual searchers "in the wild." We divide the review according to two locations in which search for audiovisual material often takes place: on the Web, and in audiovisual archives. There are important differences between web-based multimedia search engines and audiovisual archives: web-based search engines serve the general online public rather than a group of specialist users; they offer access to amateur video of varying quality as well as professionally produced material of the type curated by archives; and the search engines allow users to search on text obtained from web pages or user tags, rather than the manually created catalog descriptions that are maintained by archives.

2.2.1 Web Searchers

Searching for video has become part of daily life for many computer users. We have continuous access-on-demand to unlimited quantities of amateur and professional footage, enabled by the Web.

In one of the earliest studies of online audiovisual search, Ozmutlu et al. [118] performed a transaction log analysis of an Internet search engine. They compared multimedia queries from 2001 to queries from 1997–1999, and found that queries were changing rapidly as web content and searching techniques evolved. In particular, search sessions were becoming shorter with fewer query modifications, while queries themselves contained increasing numbers of terms. Jansen et al. [81] found that multimedia web searching was relatively complex as compared to general web search, with a longer average query length and higher use of boolean operators, in a study of transaction logs gathered in 2002. In a later study of transaction logs gathered in 2006, Tjondronegoro et al. [167] found that multimedia searches used relatively few search terms. In addition, they found multimedia searches to be generally short in duration, with more than 50% of searching sessions being less than one minute in duration. The authors used an open-source classification tool to categorize approximately 2% of queries and found that 64% of video searches are for information about people. In a comparison with logs from earlier years they found that online multimedia search had begun to shift from entertainment to other categories such as medical, sports, and technology. This shows that video search is not a static field; searcher needs are changing, though whether that change is due to changing demographics of online searchers or other factors such as changes in search engine technology is not clear.

Cunningham and Nichols [28] performed one of the few non log-based studies of web searchers for audiovisual material. They studied 48 undergraduate students and their friends. The study included both ethnographic (self)-reports and observa-

tions of search sessions. The authors used these means of analysis to investigate the motivations of their subjects when searching for audiovisual material. The authors found that, in their search for video, these users were often driven by their mood or emotional state, for example being bored and wanting to find videos to pass the time. Other motivations included locating instructional videos, and social reasons (for example, locating videos recommended by friends). They also found video search to be deeply embedded in the day-to-day lives of the users, and to typically be interspersed with a variety of daily activities.

2.2.2 Archive Users

Now we move on from the studies of audiovisual search in the completely digitalized Web environment to studies in incompletely digitized audiovisual archives. All but one of the studies reviewed here are performed in a search environment where an archivist acts as an intermediary between the archive user and the archive content. Though the communication between the archivist and the archive user was, in many cases, digital, the audiovisual material the user was searching for was not. Audiovisual archives have, in general, only recently started to digitize their content on a large scale [187], and as a result large-scale transaction log analyses such as those described for Web users have not been performed. However, a number of specialized audiovisual archives have performed analyses of their users without the aid of transaction logs, instead analyzing by hand requests made to the archives. These requests are generally natural language texts in the form of e-mails etc. The advantage of this approach is that the underlying information need is more explicitly identified than in transaction logs, where searches are limited to short keyword queries.

Audiovisual archives primarily provide access for searchers with specialized information needs, such as for example historians or television producers. Sandom and Enser [135] analyzed 1,270 information requests to 11 (mostly commercial) film archives. They found that approximately 10% of the information requests were for known-items, specifying video titles, films by particular directors, or starring particular actors. These requests were excluded from the analysis. The remaining 90% of information requests specified the desired content or subject of the footage. These were categorized according to the Panofsky-Shatford matrix [119], which classifies different types of visual content. They found that overall 68% of requests included specific content, 56% included generic content, and 2% included abstract content, where a request could contain multiple kinds of content. This was further broken down according to client type. Their final conclusion was that, in the majority of cases, the moving image information seeker was looking for audiovisual material that illustrated specific events and showed named individuals or groups of people, in particular places or on unique dates.

Hertzum [59] manually examined 275 e-mail requests to the Deutsche Film Institut, a non-commercial archive of European films. The text of the e-mail requests, which contained on average 111 words, was used to categorize them into search types: known-item retrieval, fact retrieval, subject retrieval, exploratory search, and other. He found that most requests could be classified as known-item retrieval (43%) or subject retrieval (32%) requests. A further analysis of the e-mail requests showed that 75% of the requests specified production-related attributes such as the title, director, or production year of a film. Only 38% of requests specified content, subject, or context attributes. This is in contrast to the study by Sandom and Enser [135], where 90% of the requests studied desired content or subject matter. Hertzum speculated that such discrepancies reflect differences in the requester's tasks.

Jörgensen and Jörgensen [88] analyzed 685 searches obtained from the logs of a commercial image archive. The archive did not include any audio media, and is therefore not strictly an audiovisual archive. Nevertheless, we include this study in our overview, as it is one of very few studies of commercial archives of non-textual media. The broad goal of their research was to "fill the knowledge gap concerning 'typical' search sessions [...] and to provide a general picture of searching by a particular group of image professionals." They found that unique term searches (searches for proper nouns, which refer to a specific concept, in other words, a type of navigational query) were less frequent than other studies in image archives [5, 40] had found. They also found a large amount of query modifications (62% of queries), indicating that often the information need of the searcher was not completely fulfilled by the results of the first query to the archive.

2.3 Laboratory Experiments

Now we turn from studies of searchers in their natural environment to studies of users in controlled laboratory environments. In laboratory experiments aimed at eliciting requirements for experimental video search engines, (potential) searchers are brought into a controlled setting. The impact of external influences on the experiment is minimized by keeping, as much as is possible, all variables constant, and then altering a single experimental variable. In the majority of cases such experiments are aimed at examining the performance and perception of searchers on experimental retrieval systems, rather than commercially available systems.

In order to enable performance oriented evaluations, researchers gather explicitly requested user relevance data — specifically, selected relevant videos for a set of queries [20, 32, 64, 65, 168, 184]. In addition, implicit user data in the form of logged actions may be gathered [20, 48]. It is worth noting that an important source of queries and judgments of relevant videos is the TREC Video Retrieval Evaluation

(TRECVID) [146], which is described in detail in Section 6.1.

User perceptions are taken into account by requesting that the searcher describe their experiences interacting with the retrieval system(s). Such subjective data may be gathered by means of questionnaires, interviews, and discussions [20, 32, 64, 65, 168, 184].

A problem when performing laboratory experiments is gaining access to test subjects. Presumably as a consequence of this, there is a clear divide as to the background of users participating in such experiments. The majority utilize students and research staff of academic institutions. In addition, there have been some studies of professional video searchers, and studies employing simulated users. Accordingly, we discuss each in turn below.

2.3.1 Experiments Involving Students and Researchers

When students and researchers are used as test searchers, as part of a study of audiovisual search, the research goal is generally to evaluate the effect of changing variables in experimental audiovisual retrieval systems. This is measured by examining which retrieval system helps the user find the maximum number of relevant pieces of video within a given period of time. Thus the focus here is on improving system design in terms of attained retrieval performance, rather than examining users. Such experiments are valuable for creating effective search systems, but a criticism is that performance is bound to system design, and it is difficult to draw conclusions about what techniques work outside the context of a particular system [146].

In terms of retrieval interfaces, multiple studies have demonstrated that displaying temporally related video segments, as well as video segments related to a query, improves retrieval results (e.g., [33, 56]). De Rooij and Worring [32] found that displaying even more related segments on the screen—for example, according to visual similarity—can improve video retrieval performance, though performance degrades as the number of lists on the screen becomes too large. This was confirmed by feedback from questionnaires. An example of an interface incorporating both temporally related video segments, and other related segments, is shown in Figure 2.1. Hauptmann et al. [56] investigated the effect of presenting users with results very rapidly. Though this mode of video retrieval requires absolute attention from the user, they found that the approach improved video retrieval performance. The searchers' qualitative perception of this mode of presentation was not reported.

Turning to retrieval algorithms, Hopfgartner et al. [64] found that their searchers preferred a system that provided recommendations on the basis of feedback from earlier searchers, to a system that did not provide such recommendations.

In an experiment aimed at identifying differences between types of users, Halvey and Jose [48] investigated the search behavior and perceptions of expert and novice

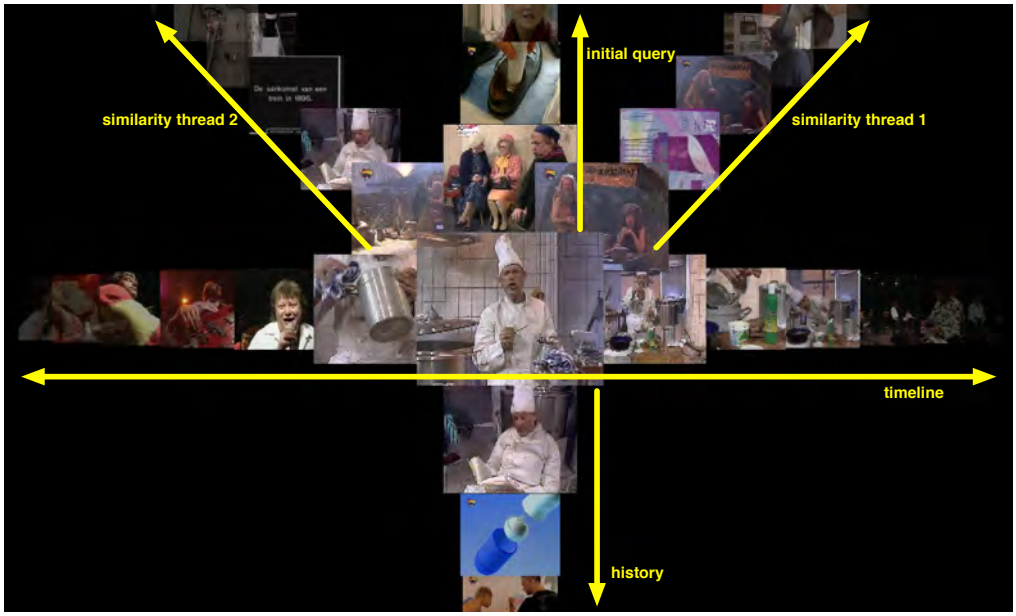


Figure 2.1: Example screenshot from a search system that displays temporally related video segments (i.e., consecutive shots in a sequence) on the horizontal axis, while video segments related to the query are displayed horizontally. Screenshot is of the Fork-Browser [34], reproduced by permission.

users on an experimental retrieval system. They found that researchers with prior experience of experimental (content-based) video retrieval systems tended to perform better in terms of retrieval performance than students with no experience with the systems. In addition, they found that expert users tended to give more negative and critical feedback than novice users, an important factor to take into account when soliciting feedback for new retrieval systems. In a large, cross-site survey, Wilkins et al. [184] found high levels of variation across users in terms of attained retrieval performance. This indicates the difficulty of quantitatively comparing search approaches on the basis of performance when there is a "user in the loop".

2.3.2 Professionals in the Lab

Christel [20] performed a laboratory experiment with a group of six government analysts, in order to analyze their search preferences. The participants were asked to perform predefined retrieval tasks using different experimental content-based video retrieval systems. These professional users preferred to use a video retrieval system with multiple query capabilities (such as query-by-example and query-by-concept)

over a video retrieval system that only allowed search on the text of transcripts. This is in contrast to their previous studies with non-professional users, described in the same paper. With these non-professional users, a single query capability (query-by-transcript) dominated. In addition, the professional users expressed disdain for tasks that did not correlate well with their daily practice; this disdain appeared to negatively influence retrieval performance. This is an interesting observation given that, in laboratory experiments, retrieval tasks are often taken from benchmarking activities and not defined on the basis of real-world use [33, 48, 56, 64, 184]. In contrast, a study of professional users in a laboratory setting that did derive search tasks from on the basis of real-world use is that of Van den Heuvel [170]. She studied transaction logs of an audiovisual archive and used them to define search tasks for eight broadcast professionals, and found that the search strategies of users were consistent across different tasks.

In Part II of this thesis we will perform quantitative retrieval experiments with experimental retrieval systems, with as our ultimate goal of improving video retrieval in the audiovisual broadcast archive setting. To cross the perceived gap between tasks taken from benchmarking activities and professional practice, we will develop retrieval tasks based on information about the daily practice within the audiovisual broadcast archive. As it is difficult to quantitatively compare search approaches with "users in the loop," we will separate our retrieval experiments from the searchers, first investigating the searchers and their information needs, and then performing retrieval experiments *automatically*, i.e., with no user involved.

2.3.3 Simulated Searchers

As laboratory studies with human searchers are expensive to perform, some researchers have explored *simulation* of users as an alternative. In a simulation, a digitized model of the searcher is used as a proxy for a real user. The advantage of such an approach is that simulated users have infinite patience, and experiments can be repeated as necessary. In addition, the researcher has near complete control over all experimental variables. However, it is not always clear whether the simulated user performs the same way as a real user, and it is not possible to gather subjective information about how a retrieval system is experienced. Here we summarize how users have been modelled in simulation experiments.

De Rooij and Worring [32] simulate an "ideal" user who issues a single query to the system, navigates through multiple result lists that are visible on the computer screen, and selects every relevant result that is encountered. This simulation is used to evaluate the potential benefit of displaying multiple result lists on the screen over displaying a single result list. The authors find that including extra result lists on the screen would give improved retrieval performance, however, in a separate eval-

uation with a group of human users, described in the same paper, the display of many result lists was indicated to be confusing, and including more result lists did not result in increased performance. This highlights a difficulty of simulation experiments; by definition a simulation model reduces the complexity of the real world [136], and there cannot incorporate all aspects of a real-world retrieval system.

Hopfgartner and Jose [63] use simulation to investigate the effectiveness of using implicitly gathered user interaction data such as video playback behavior to inform retrieval algorithms. In the simulation, different types of feedback are assigned to relevant results that are displayed in the simulated interface. Different types of feedback are shown to give different retrieval performance, but there is no relationship back to real users. In later work, Hopfgartner et al. [65] create simulated users to evaluate the performance of a faceted video retrieval system. This simulation goes beyond result browsing, and includes the formulation of new queries by adding frequently occurring keywords in clusters of results. Some simulations are informed by the logged behavior of users from a separate laboratory experiment, which results in improved retrieval performance, but no attempt is made to validate the realism of the simulation model itself.

2.4 Discussion

.

In this chapter we have reviewed studies of people searching for audiovisual material. Outside of the laboratory, studies of searchers on the Web have been dominated by the analysis of the electronic traces of user behavior recorded by search engines. Studies of searchers in audiovisual archives, on the other hand, have been restricted to small-scale studies of information requests given to archivists. Within the laboratory environment, experiments are performed under controlled conditions, and are often aimed at evaluating user performance on, and perception of, novel audiovisual search systems. When real users are not available or too expensive, simulated searchers may be employed as proxies.

Transaction log analysis allows us to quantify searcher behavior on a large scale, although it is difficult to determine the underlying information intention of the user on the basis of their search words alone. Studies in audiovisual archives have been limited to manual analyses of a requests issued to archivists. In Chapter 3 we will perform a large-scale transaction log analysis of an audiovisual broadcast archive. In this analysis we will take a step towards determining the underlying intention of the user by connecting terms in queries to terms in clicked results, and then using an audiovisual thesaurus to categorize the query terms.

Simulation allows us to study users even when we don't have users; a simulated

searcher has unlimited time and unlimited patience. However, it can be difficult to address the usefulness of simulators when they are not validated against real-world tasks. In Chapter 4 we will create simulators that attempt to reproduce, for retrieval experimentation purposes, searches and the corresponding purchases on the basis of audiovisual broadcast archive catalog data. In order to determine the validity of these simulated searches and purchases, we will compare them to the actual searches and purchases that we witness in the transaction logs.

In the review of related work in this chapter we have placed the emphasis on *searchers* rather than on *search systems*, which will be the focus of Part II. However, to ensure that evaluated retrieval tasks reflect real world tasks, observations of searchers are vital to informing task design. Therefore, in Chapter 7 we will define two evaluation collections, one of which is based on observations of searchers in the audiovisual broadcast archive. This collection will then be used later on in the thesis to perform experiments studying the possible impact of new content-based video retrieval techniques on the real world search task.

Chapter 3

Search Behavior of Media Professionals

In this chapter we present a large-scale analysis of the search behavior of media professionals in an audiovisual broadcast archive. Our study takes place at the Netherlands Institute for Sound and Vision, the national audiovisual broadcast archive that we introduced in Chapter 1, and which we will refer to in this chapter as "the archive."

We perform our study by means of a transaction log analysis, the non-intrusive user study method reviewed in Section 2.1. The transaction log analysis is enhanced by leveraging additional resources from the audiovisual broadcasting field, which can be exploited due to the specialist nature of the archive. In particular, we analyze purchase orders of audiovisual material, and we use catalog metadata and a structured audiovisual thesaurus to investigate the content of query terms.

Our study differs from previous investigations of audiovisual broadcast archive searchers in that it departs from the small-scale analysis described. It differs from analyses of online video search in that the searchers are professionals, who are searching for audiovisual material for reuse in new productions (rather than for personal consumption).

In our transaction log analysis we aim to shed light on the manner in which media professionals search through an audiovisual archive, and to gain insight into the types of content that they are searching for, thereby answering our first question,

> **RQ 1** *What kinds of content do media professionals search for, and in what manner do they search for it?*

Our question may be answered at multiple, increasingly specific, units of analysis,

as described in Section 2.1: the analysis of *sessions* containing sequences of searcher actions; the analysis of *queries* issued in the sessions, and the analysis of individual *terms* contained within the queries. In addition, there is a unit of analysis specific to our study: the purchase *orders* placed by searchers for audiovisual material. We pose four chapter-level research questions according to these units of analysis:

CRQ 1 What constitutes a typical search session at the archive? *How long does a session typically take? How many queries does a user issue in a session? How do queries develop within a session? Do sessions resulting in an order differ from sessions where no order is placed?*

CRQ 2 How are users issuing queries to the audiovisual archive? *Which search options are being used? What are commonly occurring queries?*

CRQ 3 What type of content is requested in the terms contained in queries issued to the archive? *Are terms often in the form of program titles? What is the distribution of terms across thesaurus facets? What are commonly occurring terms?*

CRQ 4 What are the characteristics of the audiovisual material that is finally ordered by the professional users? *Do users order whole programs, or fragments of programs? What is the typical duration of ordered material? Is there a preference for recent material?*

When answering these questions we will highlight the similarities and differences between the archive and other audiovisual search environments through a comparison to other transaction log studies. By delving into the characteristics of the orders made in the archive, we exploit an unusual form of information that is present in the professional archive, namely, explicit indications of the desired audiovisual content given in the form of purchase orders. While we are limited in terms of interpretations of the log data due to the lack of direct interaction with users, we are able to generalize over the large amount of data that is gathered. The results of the analysis can serve as a baseline survey for designers of systems for this type of archive, quantifying the status quo in the studied archive. Thus, the effect of alterations to the search system can be compared to the baseline.

3.1 Organizational Setting

As promised in Chapter 1, our study takes place at the Netherlands Institute for Sound and Vision, the main provider of audiovisual archive material for broadcasting companies in the Netherlands. In an international context, the archive is a representative example of the broader class of national audiovisual broadcast archives.

The archive is actively used by media professionals from a range of production studios, primarily to find material for reuse in new productions.

A core task of the archive is to manage and preserve the Dutch audiovisual heritage; this includes contextualizing the audiovisual material by manually creating textual metadata for individual audiovisual broadcasts in the form of catalog entries. Each catalog entry contains multiple fields that contain either freely-entered text or structured terms; see Figure 3.1c for an example entry as displayed in the search result interface. While the audiovisual broadcasts in the archive consist largely of television broadcasts, there are also movies, international news exchange footage, amateur footage, and internet broadcasts. There is also an audio-only collection, which consists primarily of radio broadcasts and music recordings.

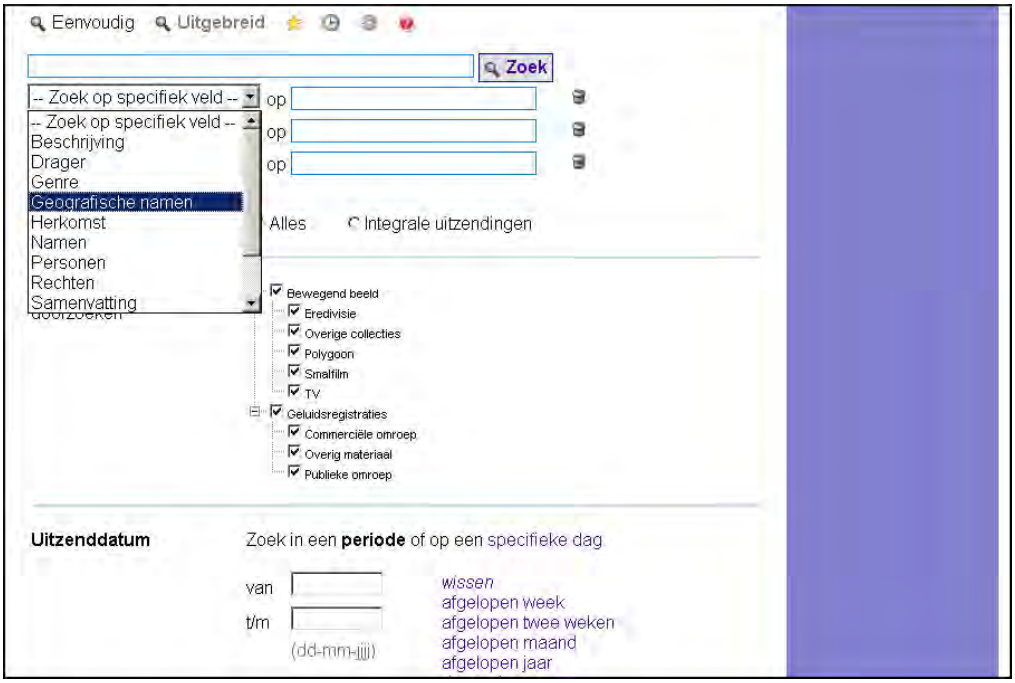
3.1.1 Users of the Archive

The archive is used primarily by media professionals, though there is some incidental use of the archive by members of the general public and others interested in the archive's content. The media professionals work for a variety of broadcasting companies, public and private, and are involved in the production of a range of programs such as documentaries and current affairs programs. The group of media professionals also includes employees of the archive's internal customer service department, who search for material on behalf of external customers [170]. Broadcast production companies typically employ one or more researchers whose sole task is to search for reusable audiovisual material. Production companies with a small team of editors usually let their editors perform the searches. In most cases, the decision on whether or not to use a candidate piece of audiovisual material in the final cut of a program is made jointly by the researchers, editors, journalists, producers, and directors. It is not uncommon for researchers to transfer their search results to another staff member who then does the final ordering for various programs. This transfer is made possible by the archive's online interface. Ordered audiovisual material is reused in many types of programs, especially news and current affairs programs, talk shows based on daily news, late-night shows, and programs about history. In addition, it is sometimes used for other purposes; for example, to populate online platforms and exhibitions.

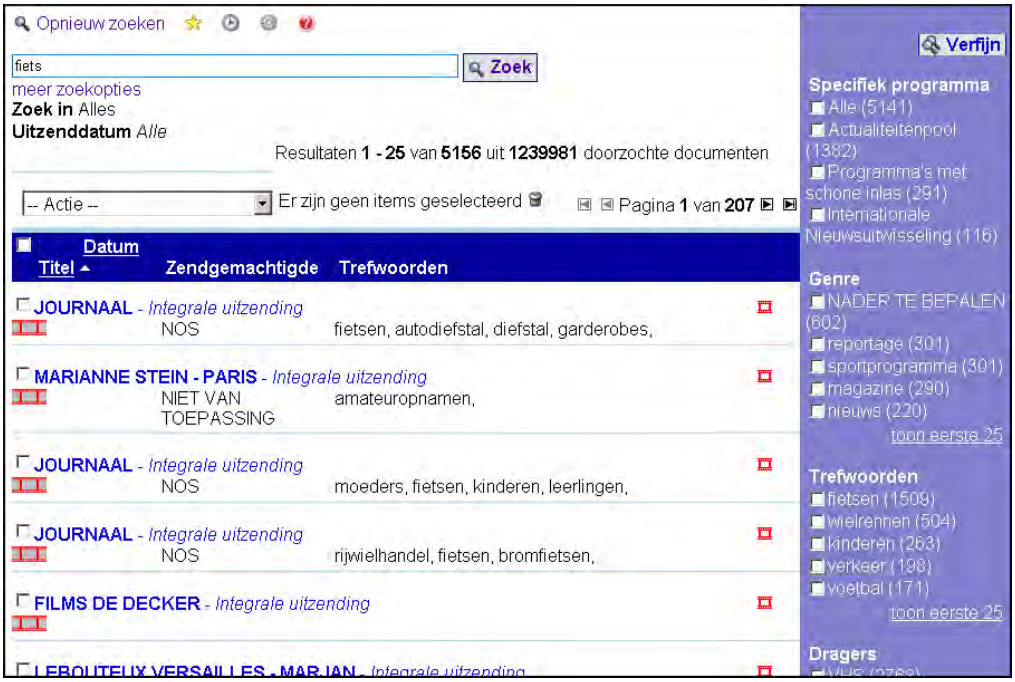
3.1.2 Finding and Ordering Audiovisual Material

To support users in sourcing audiovisual material, the archive has developed a retrieval system, iMMix[1]. Screen shots of the retrieval interface are shown in Fig-

[1] A publicly available version of the iMMix search system can be accessed online at `http://zoeken.beeldengeluid.nl/`.

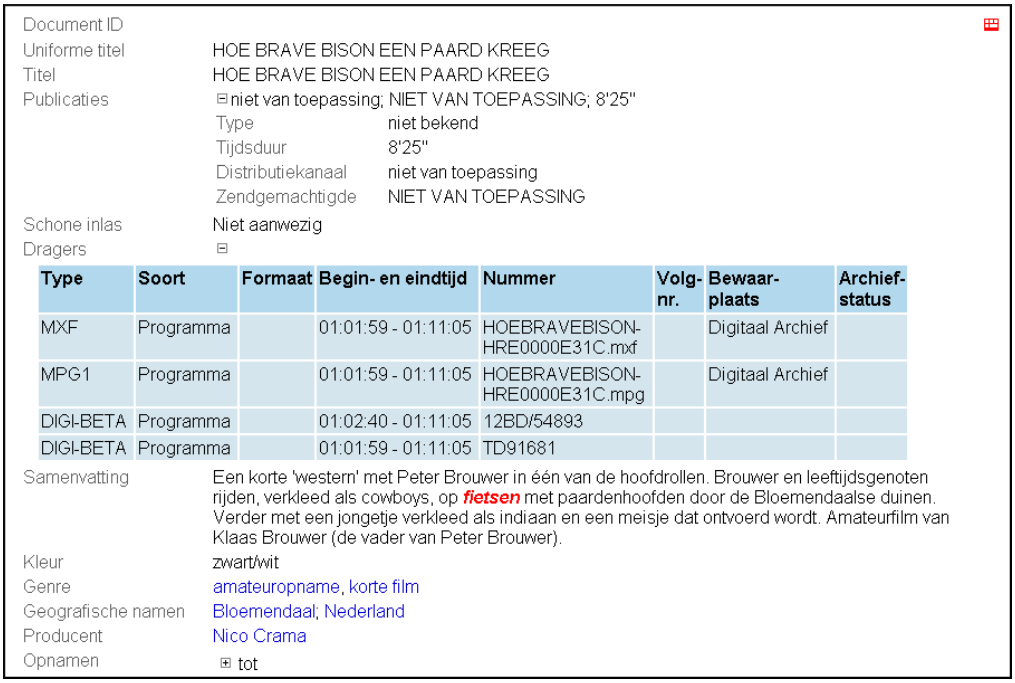


(a) Advanced search form

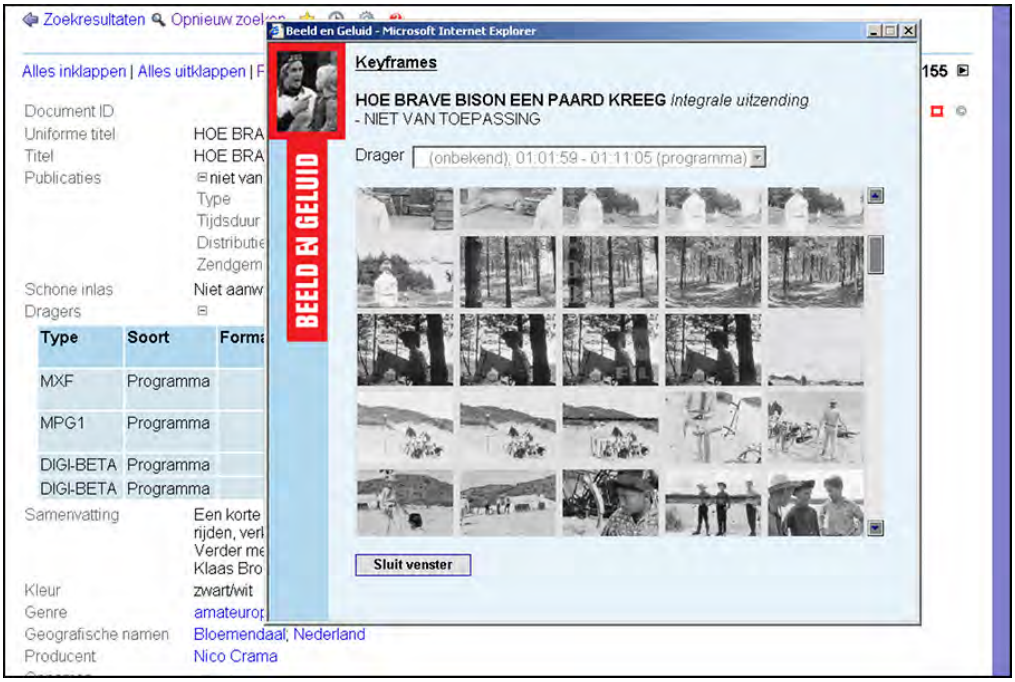


(b) Results delivered by search, in response to the query “fiets” (bicycle)

Figure 3.1: Screenshots of the interface of the archive’s search engine.



(c) Item details of clicked result for the query “fiets” (bicycle)



(d) Keyframe preview of video content for the clicked result from 3.1c

Figure 3.1: Screenshots of the interface of the archive’s search engine (continued from previous page.)

ure 3.1. The iMMix interface contains a simple search page and an advanced search page; the latter is shown in Figure 3.1a. The search pages offer the following three search options:

Keyword search Search of freely-entered text, returning results according to the textual content of all catalog fields

Date filtering Return results broadcast on a specific date or within a set period

Advanced search Search within specific fields from the audiovisual catalog

After the search results have been displayed, they can be refined by using a list of suggested filtering terms displayed on the right-hand side of the result page; see Figure 3.1b. These suggested terms are based on the content of the different fields of returned results, allowing the user to refine her search without manually typing in values in the advanced search field. The user can navigate to several types of detailed result views: the catalog entry for the result as shown in Figure 3.1c, keyframes from the result as shown in Figure 3.1d, or a video or audio preview of the result in a media player. The latter two views are only available if a result is digitally available. Once a user has decided to purchase a result, or a fragment of footage from the results, the material can be ordered using an online form. Ordered video items are paid for on a per-second basis. Radio can only be ordered in chunks of at least one hour. Rates vary per copyright owner and per user subscription type, and some items can be obtained free of charge. Through iMMix, video material can be ordered as a fragment or as an entire program. Reasons to order a fragment rather than an entire program include: increased efficiency during the editing process, budget limitations, and copyright restrictions that mean that only some parts of a program may be reused.

3.1.3 Audiovisual Thesaurus

Search through audiovisual material in the archive is based on manually created catalog entries. These entries are created by the archival staff, and include both free text as well as structured terms contained in a specialized audiovisual thesaurus. The thesaurus is called the *GTAA* — the Dutch acronym for "Common Thesaurus for Audiovisual Archives" [12]. The GTAA thesaurus is continually evolving, with new terms being added subject to an internal review process. As of June 2009 it contained approximately 160,000 terms. The thesaurus terms are organized in six facets:

Location Either the place(s) where audiovisual material was recorded, or the place(s) mentioned or seen in a broadcast.

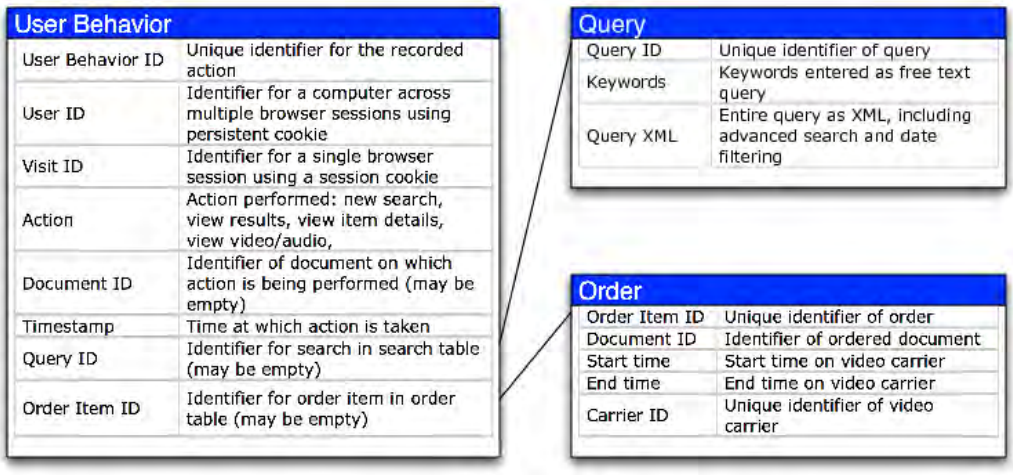


Figure 3.2: Overview of the utilized tables and fields from the transaction logs. All actions are recorded in the *User Behavior* table. Additional information about queries and orders issued to the system is recorded in the Query and Order tables.

Person The proper names of people who are either seen or heard in a broadcast, or who are the subject of a broadcast

Name The proper names of entities that are not people or locations, that are either seen or heard in a broadcast, or who are the subject of a broadcast. Includes named organizations, groups, bands, periods, events, etc.

Program Maker A person involved in creating a broadcast.

Genre The genre of a broadcast, such as *news* or *satire*.

Subject Keywords describing what a broadcast is about and what can be seen in it.

The iMMix search engine indexes the terms in these structured thesaurus facets, and also the terms in other, unstructured fields such as free text descriptions and summaries of the videos, and technical information such as the format of the video data.

3.2 Experimental Design

3.2.1 Data Collection

Transaction logs from the archive's online search and purchasing system were collected between November 18, 2008 and May 15, 2009. The logs were recorded using

an in-house system tailored to the archive's online interface. The logs were stored in a database with multiple tables. For this study we utilized three main tables, illustrated in Figure 3.2: the *User Behavior* table, which recorded information allowing users to be connected to actions performed on the website during a visit, the *Query* table, recording information about the queries issued to the system, and the *Order* table, recording information about orders placed through the system. As is the case with many online systems, the archive's search interface is periodically accessed by automated scripts. This phenomenon is typified by a large number of repeats of the same query. To minimize the impact of these non-human interactions on the query log analysis, traces left by users issuing over 80 identical queries on a single day were removed from the logs. In addition, log entries generated by the archive's search system development team were also automatically removed. In total, the logs contained 290,429 queries after cleaning.

The transaction logs often reference documents contained in the archive's collection. In order to further leverage the log information, we obtained a dump of the catalog descriptions maintained by the archive on June 19, 2009. The size of the catalog at that time was approximately 700,000 unique indexed program entries.

3.2.2 Definitions

We define five key units that will play a role in our analysis below: the three units common in transaction log analysis of *session*, *query*, *term*; and two units specific to this study, *facet* and *order*. The specialized knowledge sources available within the archive allowed (and motivated) us to develop the last two units of analysis.

Session A portion of a user's visit spent searching for items addressing the same subject or topic, as identified by overlapping words between queries [82]. Session boundaries within a single visit are assigned when a query has no words in common with the previous query. Some studies have included a session timeout to identify session boundaries. We do not follow their example as in this study users routinely view audiovisual material during the search process; this can lead to periods of inactivity even though the user is still fully engaged in the search process.

Query A single information request issued to the archive, using any of the three previously described search options available in the interface: (1) *keyword search*, (2) *date filtering*, and (3) *advanced search*. A query may also be empty, for example when a user clicks the search button without entering information in any of the search boxes.

Term A single *thesaurus term* or *title term* contained in a query, identified by matching phrases and words from the query to thesaurus entries and titles from

Table 3.1: Overall transaction log statistics.

Description	**#**
Search sessions	139,139
Queries	
- Empty	15,675
- Non-Empty	274,754
Matched Terms	
- Title Terms	72,620
- Thesaurus Terms	83,232
Orders	27,246

clicked results. For example, if a user issues a query for "obama white house" and clicks a result containing the thesaurus entry "barack obama," then the query is considered to contain the thesaurus term "barack obama." This content-based coding approach is similar to that adopted by Jörgensen and Jörgensen [88] in that content-bearing concepts are identified as terms, but differs in that the coding process is done automatically. As a consequence, term identification can be done over large numbers of queries, but not all terms can be identified. A detailed description of the term matching approach and its evaluation is given in Appendix A.

Facet The predefined category for a thesaurus term, as assigned within the GTAA in-house audiovisual thesaurus described earlier. This is either *Location*, *Person*, *Name*, *Program Maker*, *Genre*, or *Subject*.

Order A purchase request for audiovisual material issued to the archive. An order may be for an entire broadcast or a subsection of a broadcast. This is an unusual form of information that is available due to the dedicated professional nature of the archive.

3.3 Transaction Log Analysis

To give an overview of our transaction log analysis, Table 3.1 gives general statistics about the data collected in terms of the number of queries issued, query terms matched to the audiovisual thesaurus, search sessions, and the number of orders made. The data was collected over a period of 178 days, giving an average of 782 sessions per day. However, sessions are not equally divided over the days of the week, and a session may last longer than a day, as we shall see below.

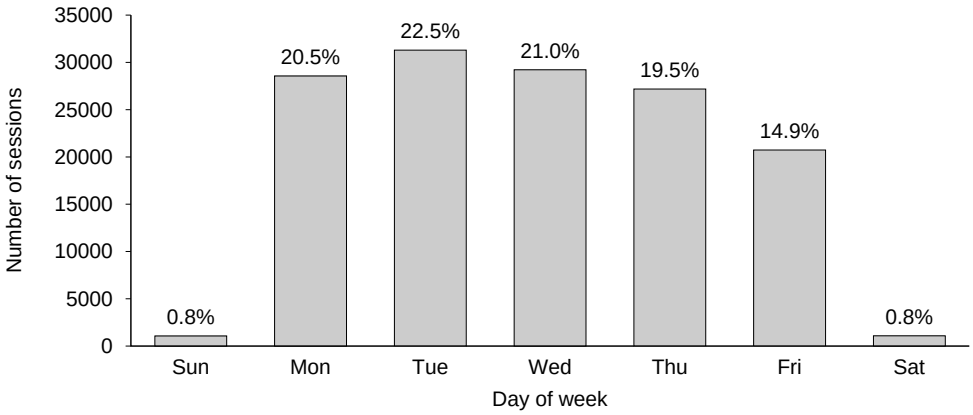


Figure 3.3: Distribution of sessions across days of the week. Most activity takes place on weekdays.

3.3.1 Session-level Analysis

Distribution over the Week Figure 3.3 illustrates the distribution of sessions across the days of the week. As in [88], we find that the majority of sessions occur on Monday through Friday, reflecting the professional nature of the archive use. Moreover, fewer sessions occur on Fridays, reflecting an ongoing societal development in the Netherlands: increasing numbers of people have four day working weeks (with the weekend starting on Friday).

Session Duration The session statistics shown in Table 3.2 provide us with insights into the characteristic behavior of users of the archive. However, before we analyse the implications of the data, we briefly discuss its statistical representation.

As is customary in transaction log analyses, Table 3.2 gives mean values for variables such as queries per session, result views per session, and order duration. However, as is shown by the standard deviation values — that are higher than their corresponding means, and indicate that negative measurements are normal — there is excessive variability in the data. This corresponds to the findings of Jansen et al. [82], who observed average session duration lengths to be unstable while the large percentages of short sessions remained more stable.

In our own case we found that variability of search sessions was caused by some users remaining logged into their browser, leaving their workstation for one or more nights, and returning to complete the search session that they previously initiated. The 2008 Christmas holidays were a contributing factor here, with users continuing to be logged on over the holidays, and the longest session lasting more than

Table 3.2: Session-level statistics over the entire transaction log collection. Statistics are given over all sessions, and also according to whether an order was placed during a session or not. Note that the standard deviation values are larger than their corresponding mean values, motivating our use of median measurements.

			Session resulted in order?	
Variable	**Measure**	**All sessions**	*Yes*	*No*
Total sessions	Count	*139,139*	15,403 (11%)	123,736 (89%)
Queries	Median	*1.0*	1.0	1.0
	Mean	*2.0*	2.6	1.9
	σ	*2.4*	3.3	2.3
Result views	Median	*1.0*	1.0	1.0
	Mean	*2.2*	2.1	2.2
	σ	*4.5*	4.4	4.5
Orders	Median	*0.0*	1.0	0.0
	Mean	*0.2*	1.8	0.0
	σ	*1.0*	2.5	0.0
Duration	Median	*00:00:59*	00:07:47	00:00:43
(hh:mm:ss)	Mean	*00:17:42*	00:45:31	00:14:09
	σ	*03:59:52*	05:19:04	03:47:32

800 hours (approximately 1 month). Consequently, the average session duration of approximately 18 minutes has a very high standard deviation of four hours. This suggests that the median duration, indicating that 50% of sessions have a duration of 59 seconds or less, is a more representative indicator of session length. These results are similar to the median session durations of online multimedia search engines reported by Tjondronegoro et al. [167], who found that 50% of the searches took less than a minute. Session durations reported at a professional image archive, on the other hand, were longer, with a median session time of five to six minutes [88]. In the rest of our discussion, we will discuss median values rather than mean values.

Sessions Resulting in Orders The discrepancy between session duration in the audiovisual archive and the session duration in the professional image archive can perhaps be explained by considering the sessions that resulted in an order separately from those where no order for audiovisual material was placed. The 11.1% of sessions that resulted in an order exhibit a much longer median duration of over 7 minutes, and are more consistent with the results of the professional image archive [88]. The increased duration for sessions resulting in an order is not accompanied by a similarly large increase in the number of queries issued and results viewed. This suggests that the extra time spent is primarily spent on other tasks such as reviewing audiovisual content or placing the order. We will examine this in more detail

below, in the analysis of orders placed to the system.

Number of Queries per Session Figure 3.4 gives more insight into the numbers of queries issued per session. As implied by the median statistics in the previous paragraph, the majority of sessions contain one query. A number of sessions contain no query at all: these are sessions where, for example, a user navigated to the search page but did not issue a query.

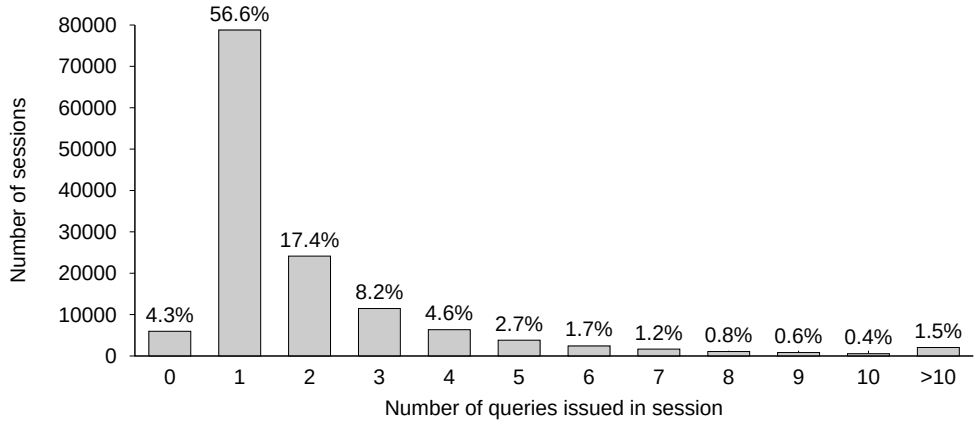


Figure 3.4: Overview of the numbers of queries issued per session.

Query Modification We also look at the development of queries within a session, examining the number of queries used in each session, and how successive queries differed from previous queries in the same session. The results are shown in Table 3.3. Similarly to Jansen et al. [80], we classify our queries as *initial*, *modified*, or *repeat*. An *initial* query is the first query issued in a session, and therefore represents the number of sessions where at least one non-empty query was issued. A *modified* query is a query that is different from the previous query. A *repeat* query is a query that is identical to the previous query: this may be due to the user retyping the same query words, but can also be caused by the user reloading the search page. The transaction logs do not allow us to distinguish between these underlying causes; thus, it is not possible to interpret these queries meaningfully. The proportion of query modifications is relatively small as compared to [88], where 61% of queries were found to be modified.

Lessons Learned The transaction logs contain a large number of short sessions, and also contain a small number of very long sessions. These can sometimes last for

Table 3.3: Number of initial, modified, and repeat queries recorded in the search logs.

Modification type	**#**	**%**
Initial	132,289	48%
Modified	95,274	35%
Repeat	47,191	17%

Table 3.4: Breakdown of different search options used in the logged non-empty queries. Multiple search options may be employed in a single query.

Functionality	**#**	**%**
Keyword search	264,270	96%
Date filtering	62,647	23%
Advanced search	23,332	9%

weeks, as users may not log out of their browser while away from their workstation. Though most sessions are very short, sessions resulting in an order are generally much longer than sessions that do not result in an order. The longer session duration is not accompanied by a corresponding increase in queries or result views. Most sessions consist of a single query.

3.3.2 Query-level Analysis

Now we turn to an examination of the queries issued to the archive, studying in particular the utilization of the different search options afforded by the archive's interface, and summarizing the most frequently issued queries.

Search Option Utilization As we described previously, the search interface provides users with multiple options for search: keyword search, date filtering, and advanced search of specific catalog metadata fields. A query may be specified using any or all of these search options. Table 3.4 gives a breakdown of the utilization of the available search options over the non-empty queries in the transaction logs. The clear majority of queries contain a keyword search. Interestingly, almost a quarter of the queries in the logs incorporate a date filter. The advanced search option, which allows users to search the catalog for thesaurus terms used to annotate broadcasts as well as for specific technical properties (e.g., copyright holder, physical format), was the least used of the three search options.

Frequent Queries Table 3.5 separately lists the most frequently occurring queries for each of the three search options offered by the archive, with English language descriptions where appropriate. Of the top 20 most frequent keyword searches, 17

Table 3.5: Most common searches, split according to search option into keyword search, advanced search, and date filter.

Keyword search	**Description**	**#**	**%**
journaal	Title of news series	3933	1.49%
nova	Title of current affairs series	1741	0.66%
eenvandaag	Title of current affairs series	1340	0.51%
netwerk	Title of current affairs series	1103	0.42%
journaal 20	Title of news series	1052	0.40%
brandpunt	Title of current affairs series	807	0.31%
jeugdjournaal	Title of news series	778	0.29%
nos journaal	Title of news series	674	0.26%
het klokhuis	Title of educational series	517	0.20%
de wereld draait door	Title of talk show	439	0.17%
nos	Name of broadcasting company	416	0.16%
polygoon	Title of news series	396	0.15%
videotheek	Physical media location	393	0.15%
andere tijden	Title of documentary series	369	0.14%
twee vandaag	Title of current affairs series	339	0.13%
kruispunt	Title of current affairs series	337	0.13%
obama	Person name	317	0.12%
klokhuis	Title of educational series	284	0.11%
20 uur journaal	Title of news series	284	0.11%
20	Title of news series	265	0.10%

(a) Most frequently occurring keyword searches accompanied by an English language description. Percentages are with respect to the total of 264,270 keyword searches.

Advanced search	**#**	**%**
format:mxf	5337	22.87%
copyright holder:nos	800	3.43%
genre:documentaire	772	3.31%
format:digi-beta	728	3.12%
copyright holder:vara	451	1.93%
copyright holder:ncrv	447	1.92%
copyright holder:vpro	398	1.71%
copyright holder:tros	383	1.64%
format:mpg1	380	1.63%
copyright holder:none	378	1.62%
format:vhs	354	1.52%
copyright holder:teleac	331	1.42%
copyright holder:eo	241	1.03%
copyright holder:kro	229	0.98%
copyright holder:nps	222	0.95%
copyright holder:avro	205	0.88%
format:mxf, copyright:ncrv	188	0.81%
genre:journaal	147	0.63%
format:wmv	140	0.60%
format:mpg1, format:mxf	127	0.54%

(b) Most frequently occurring advanced searches, given in the format *field name:value*. Percentages are with respect to the total of 23,332 advanced searches.

Start Date	**End Date**	**#**	**%**
2000-01-01	2009-12-31	1083	1.73%
1980-01-01	1989-12-31	410	0.65%
1990-01-01	1999-12-31	309	0.49%
1970-01-01	1979-12-31	308	0.49%
1960-01-01	1969-12-31	276	0.44%
2009-03-12	2009-03-12	201	0.32%
2009-02-25	2009-02-25	187	0.30%
2009-01-26	2009-01-26	174	0.28%
2009-01-20	2009-01-20	152	0.24%
2008-11-05	2008-11-05	147	0.23%
2009-01-09	2009-01-09	147	0.23%
2008-01-01	2008-12-31	142	0.23%
1950-01-01	1959-12-31	141	0.23%
2008-11-26	2008-11-26	138	0.22%
2009-03-09	2009-03-09	138	0.22%
2008-11-18	2008-11-18	132	0.21%
2009-01-18	2009-01-18	132	0.21%
2009-02-17	2009-02-17	129	0.21%
2008-11-25	2008-11-25	125	0.20%
2008-12-31	2008-12-31	124	0.20%

(c) Most frequently occurring date filters used in queries. Percentages are with respect to the total of 62,647 date filtering searches.

consist of the title or partial title of an audiovisual broadcast. Furthermore, the titles are mostly of Dutch news programs, or Dutch current affairs shows. However, the frequently occurring list accounts for only 5.6% of all the keyword searches in the archive. It is not clear from Table 3.5 whether the long tail of searches is also dominated by the titles of Dutch news and current affairs broadcasts. We will address this in Section 3.3.3 with a content analysis of the terms contained in the keyword searches.

Advanced searches are commonly used to specify the format of the results to be returned, with almost a quarter of all advanced queries (of which the most frequent are shown in Table 3.5) specifying that only audiovisual documents available in the digital *Material eXchange Format* (MXF) should be returned.

The date filter is commonly used to specify date ranges of a decade, which account for the five most frequent date ranges in Table 3.5, or otherwise a single day.

Lessons Learned Almost all queries issued to the archive contain keyword search terms, and an analysis of the most frequent searches indicates that common keyword searches tend to be derived from broadcast titles, especially news and current affairs broadcasts. The date filter is an often-used function, with results being filtered by date in almost a quarter of all queries. Advanced search appears to be used primarily for specifying technical characteristics of the desired results in terms of format and digital rights. These determine, among other things, the availability and the cost of reusing the audiovisual material returned by the system. While advanced search offers the opportunity to specify terms contained in the audiovisual thesaurus, this option is not frequently used. Instead, these types of terms are often entered in the keyword search field, as we will see in the following section.

3.3.3 Term-level Analysis

We now turn to an analysis of the query terms contained in the free text searches. We add structure to the query terms as previously described, by matching terms from the user query to titles and thesaurus entries in the results clicked by the user. The matched terms will be referred to as *title term* and *thesaurus term*, respectively. Of the 203,685 queries where users clicked a result during the session, 68,520 (34%) contained at least one title term, and 72,110 (35%) contained at least one thesaurus term. A total of 71,252 (35%) queries contained no title term or thesaurus term. The manual analysis of a subsample of queries showed that users sometimes type dates and program codes directly into the keyword search field; further detail is given in Appendix A.

Table 3.6: Most common terms occurring in keyword search, with title terms and thesaurus terms specified separately. An English language description is provided for each term.

Title Term	**Description**	**#**	**%**
journaal	News series	9087	12.5%
nova	Current affairs series	2568	3.5%
netwerk	Current affairs series	1597	2.2%
eenvandaag	Current affairs series	1440	2.0%
het klokhuis	Educational series	1318	1.8%
jeugdjournaal	News series	951	1.3%
brandpunt	Current affairs series	834	1.1%
pauw & witteman	Talk show	820	1.1%
twee vandaag	Current affairs series	779	1.1%
de wereld draait door	Talk show	728	1.0%
andere tijden	Documentary series	593	0.8%
zembla	Documentary series	444	0.6%
het zandkasteel	Children's series	409	0.6%
studio sport	Sports news series	397	0.5%
kruispunt	Current affairs series	373	0.5%
pinkpop	Documentary	367	0.5%
goedemorgen nederland	Morning show	326	0.4%
man bijt hond	Magazine show	325	0.4%
het uur van de wolf	Documentary series	294	0.4%
tv show	Variety show	292	0.4%

(a) Most frequently matched title terms. Percentages are with respect to the total of 72,620 matched title terms.

Thesaurus Term	**Description**	**Facet**	**#**	**%**
vs	U.S.	Location	1002	1.2%
nederland	The Netherlands	Location	941	1.1%
nieuws	News	Genre	695	0.8%
obama, barack	U.S. President	Person	663	0.8%
amsterdam	Dutch capital	Location	610	0.7%
wilders, geert	Dutch politician	Person	531	0.6%
irak	Iraq	Location	515	0.6%
voetbal	Soccer	Subject	460	0.6%
beatrix (koningin nederland)	Queen of Netherlands	Person	414	0.5%
bos, wouter	Dutch politician	Person	407	0.5%
bush, george (jr.)	U.S. President	Person	371	0.4%
willem-alexander (prins nederland)	Prince of Netherlands	Person	368	0.4%
kinderen	Children	Subject	342	0.4%
afghanistan	Afghanistan	Location	330	0.4%
fortuyn, pim	Dutch politician	Person	279	0.3%
ajax	Dutch soccer team	Name	272	0.3%
rotterdam	Dutch city	Location	261	0.3%
vrouwen	Women	Subject	260	0.3%
balkenende, jan peter	Dutch prime minister	Person	257	0.3%
israël	Israel	Location	251	0.3%

(b) Most frequently matched thesaurus terms. Percentages are with respect to the total of 83,232 matched thesaurus terms. The facet in which each term appears is also indicated.

Title Terms The majority of the twenty most frequent title terms, shown in Table 3.6a, are the titles of news and current affairs programs. Many of the title terms in the top twenty are identical to frequent keyword searches in Table 3.5a. In addition to news and current affairs series, the list also includes television programs such as talk shows, educational programs, and documentaries. The list gives an indication of the types of programs that are often searched for by media professionals at the archive, but gives few clues as to what kind of content they desire. News broadcasts, documentaries, talk shows and current affairs may cover a wide range of subjects and entities.

Thesaurus Terms Table 3.6b shows the twenty most frequent thesaurus terms. The list contains a variety of locations, people, and subjects. Popular locations include the United States, the Netherlands, and Dutch cities, while popular people include political figures from the Netherlands and the United States, as well as Dutch royalty. The most frequently searched for general subject is *soccer*, a popular sport in the Netherlands. Correspondingly, *Ajax*, the name of one of the dominant Dutch soccer clubs, is the most frequent non-person and non-location name. Similar lists of frequently used search terms have been presented by Jansen et al. [80], Jörgensen and Jörgensen [88], and Tjondronegoro et al. [167]. However, a direct comparison is complicated by the fact that we include multi-word terms, while Jansen et al. [80] define a term as "any series of characters bounded by white space." Both Jansen et al. [80] and Tjondronegoro et al. [167] examine queries on the Web; they contain mostly adult terms. Tjondronegoro et al. [167] do observe that next to the adult material, names of people and song titles are frequently sought, which is in agreement with our findings. Jörgensen and Jörgensen [88] find frequent search terms such as *woman*, *beach*, *computer* and *man* in their study of search in a professional image archive. It is striking that these terms are very general as compared to the majority of frequent terms that occur in audiovisual archive logs that we have analyzed here. Though the list does include general terms such as *women* and *children*, these are not in the majority.

Lessons Learned The frequent use of broadcast titles as search terms implies that users are performing a large number of navigational, or known–item, searches, as described in Chapter 2. If we take the use of title terms in a query to be indicative of known item search, the results are consistent with the findings of the early manual examination of email requests to a film archive by Hertzum [59], who found that 43% of the requests submitted to the archive were for known items. Many user queries contain terms that can be matched to audiovisual thesaurus entries, allowing us to investigate what types of content users are searching for within broadcasts.

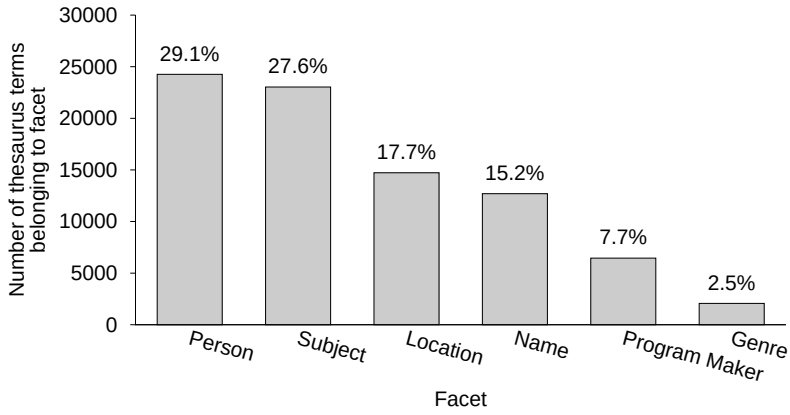


Figure 3.5: Distribution of query terms matched to the GTAA thesaurus across the six different facets contained in the thesaurus. Percentages are with respect to the total of 83,232 thesaurus terms.

Popular thesaurus terms include countries and cities, the *news* genre, and the names of political figures.

3.3.4 Facet-level Analysis

The frequently occurring terms in Table 3.6 provide us with an impression of the types of audiovisual terms users frequently search for, however the majority of thesaurus terms occur only a few times. This "long tail" phenomenon has been noted by many, including Jansen et al. [80]. The semantic, faceted, structure of the thesaurus can be used to abstract over the thesaurus terms and thereby also include infrequently used terms in the analysis.

Frequencies of Facets Figure 3.5 shows the cumulative frequency of each of the six thesaurus facets over all of the thesaurus terms identified in user queries. The *Person Name* and *Subject* facets were most frequently identified, together accounting for more than half of the thesaurus terms. They are followed by the *Name*, *Location*, and *Program Maker* facets. Overall *Genre* was the least frequently identified facet. The manual evaluation of the term identification procedure in Appendix A indicates that locations and genres are easier to match than terms from other facets, and therefore these may be somewhat overrepresented in the analysis. Nevertheless, it is apparent that the proper names of people, places, program makers, and other proper names account for a large proportion of the thesaurus terms identified in the queries.

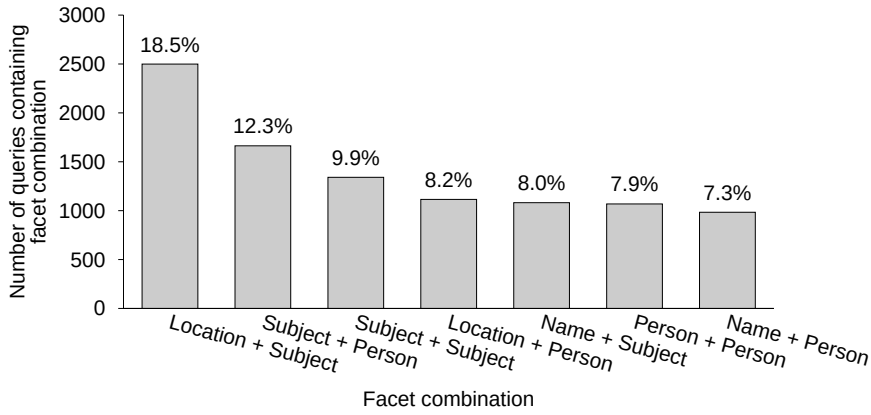


Figure 3.6: The seven most frequently co-occurring combinations of facets when considering queries containing multiple thesaurus terms. Percentages are with respect to the 13,532 queries that contain multiple matched thesaurus terms. The most common combination of facets in a query is that of *Location* and *Subject*.

Facet Combination User queries may contain multiple matched thesaurus term, and in total 13,532 queries contained more than one thesaurus term. Figure 3.6 itemizes the most frequently co-occurring facets in queries. The *Subject* facet appears frequently in combination with other facets — 49% of the combinations consist of a *Subject* and either a proper name (*Location*, *Person*, or *Name*) or another *Subject*). An example of such a combination is a query consisting of the keywords "forest fire california," which is matched to the *Subject* "forest fires" and the *Location* "california."

Lessons Learned Proper names, of people, places, events, and more, are the most common type of thesaurus term identified in queries. Together these account for 70% of the identified thesaurus terms issued to the archive. Approximately 28% of the identified terms are for more general subject-related thesaurus terms.

3.3.5 Order-level Analysis

We turn to the orders issued to the archive by media professionals, examining their age, duration, and the amount of time typically taken to place an order using the interface shown earlier in Figure 3.1.

Order Age As is clear from the analysis so far, a primary function of the archive is to supply media professionals with reusable audiovisual material for new produc-

tions. Users with an account can order and purchase material from the archive. The transaction logs of the archive in this study are exceptional in that they contain detailed information about orders of audiovisual material placed through the online interface. In addition, the orders can be connected to catalog information about the associated audiovisual broadcasts. This allows us to determine, for example, the age of the audiovisual material relative to the date of purchase. Figure 3.7 shows that the majority of ordered items (83%) were broadcast more than a month before the date of purchase; 54% of ordered items were broadcast no more than a year before the date of purchase.

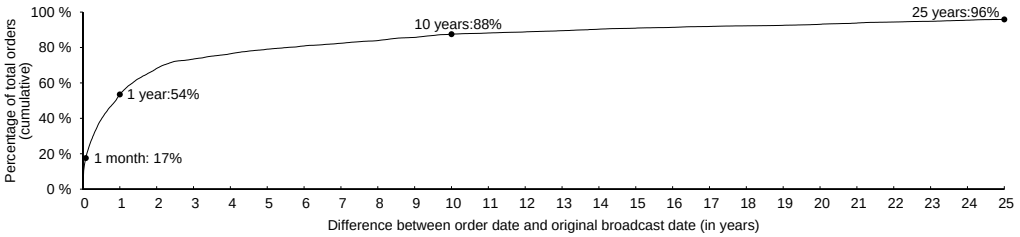


Figure 3.7: Cumulative graph showing the elapsed time since the original broadcast date, at time of ordering. The majority of purchased multimedia items was ordered less than a year after the original broadcast.

Order Units To provide a more detailed level of insight into individual orders, we categorize them according to the unit of purchase. The unit of material ordered by the professional falls into one of four categories:

Broadcast The user orders an entire broadcast.

Story The user orders a subsection of the broadcast that has been predefined by the archive catalogers as a cohesive unit.

Fragment The user orders a subsection of the broadcast that has been dynamically defined by the user, typically by using the preview or keyframe view function.

Unknown The unit of purchase cannot be determined due to missing or incomplete data.

Table 3.8 gives an overview of order statistics in terms of the number of orders, the time taken to order an item, the length of the final order, and the age of the ordered item, broken down according to the unit of purchase.

Time to Order At the median point, it takes users over twice as long to place an order for a fragment as it does for them to place an order for an entire broadcast. We

Table 3.7: Order statistics, over all orders, and separately over individual units of orders. 363 (1%) of the orders could not be classified; these orders have not been included in the table.

			Order unit		
Variable	**Measure**	**All orders**	*Broadcast*	*Story*	*Fragment*
Total orders	Count	27,246	8,953 (33%)	4,611 (17%)	13,329 (49%)
Time to order	Median	00:06:38	00:03:44	00:05:51	00:09:27
(hh:mm:ss)	Mean	00:36:34	00:50:13	00:22:58	00:32:36
	σ	04:17:49	06:56:42	00:55:24	02:13:40
Order length	Median	00:03:45	00:25:03	00:02:04	00:02:00
(hh:mm:ss)	Mean	00:12:51	00:29:41	00:03:57	00:04:40
	σ	00:29:33	00:22:41	00:07:57	00:33:27
Order age	Median	00-10-18	01-01-25	00-05-02	00-11-28
(yy-MM-dd)	Mean	04-01-21	04-02-05	01-05-05	04-11-28
	σ	07-10-29	07-11-01	03-11-13	08-07-18

attribute this to the need for a user placing a fragment order to manually review the video content to determine which portion to order. This is a time-consuming process.

Order Length There is a strong demand for short pieces of audiovisual material, with 66.7% of the material ordered being for either predefined stories or manually defined fragments, both of which tend to be quite short (approximately two minutes). The duration of ordered items is further broken down in Figure 3.8, which shows fragment orders peaking at between zero and one minute long, story orders peaking at between one and two minutes long, and broadcast orders peaking at 24 to 25 minutes in length — a common duration for a Dutch television broadcast. There are also a small number of very short broadcasts, less than five minutes in duration, shown in Figure 3.8. This is due to users ordering short complete items such as amateur recordings or international news exchange footage.

Video vs. Audio As shown in Table 3.8, nearly all orders were for video material rather than audio material. Audio results are more frequently clicked than purchased.

Lessons Learned With news and current affairs programs being created daily, one might expect the orders for audiovisual material to be dominated by very recent

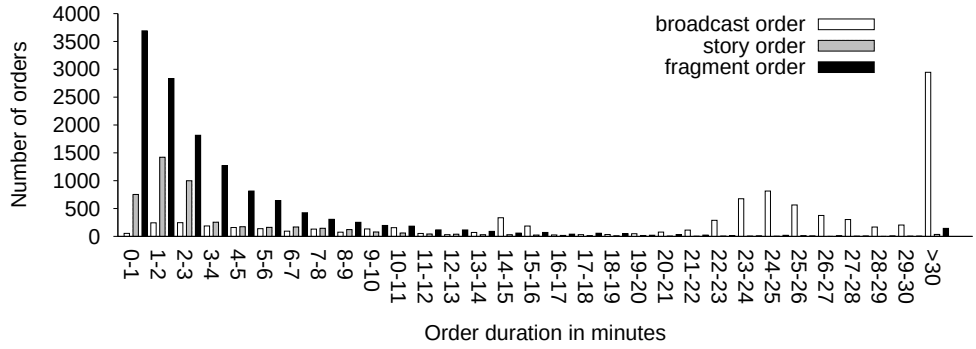


Figure 3.8: Distribution of duration of fragment orders, in minutes. Note that the last column combines all orders with a duration of 30 minutes or more.

Table 3.8: Result clicks and ordered items divided according to multimedia type: video and radio.

	Result clicks		**Ordered items**	
Multimedia type	#	%	#	%
video	287,291	91%	26,618	98%
audio	24,646	8%	262	1%
unknown	4,409	1%	356	1%

material. However, the order analysis shows that this is not the case. A little under half of the orders are for footage broadcast more than a year before the order date, and 12% of orders are for material that is over 10 years old. Users who manually specify the length of the audiovisual fragment they want to order take more than four times longer to place an order than those who simply purchase an entire broadcast. Manually specified fragments are short, often under 60 seconds in length. Internal interviews done by the archive indicate that this is often due to budgetary reasons: users pay by the second for purchased material, depending on the copyright owner. Though the archive contains both video and audio material, almost all orders placed in the archive are for video material.

3.4 Conclusion

In this chapter we have presented a descriptive analysis of transaction logs from an audiovisual broadcast archive. The analysis was structured around four chapter-level research questions, the answers to which we summarize here and follow with a discussion.

3.4.1 Answers to Research Questions

With respect to **CRQ 1**, *what characterizes a typical session at the archive?*, we found the typical session to be short, with over half of the sessions under a minute in duration. In general, there were also few queries and result views in a session, with a median value of one query issued and one result viewed. However, we found that sessions resulting in orders had a considerably longer duration, with over half of the sessions taking more than seven minutes, but no increase in terms of the number of queries issued and results viewed.

So, given that sessions resulting in order have, at the median point, as many queries and result views as sessions that do not result in an order, why do they take so much longer? This brings us to **CRQ 4**, *What are the characteristics of the audiovisual material that is ordered by the professional users?*. We isolated the orders placed to the archive and analyzed the median amount of time users took from beginning a session to finally placing an order. Here we found that users took two and a half times longer to order a fragment of audiovisual material than to order an entire program. This implies that manually reviewing the video material is a cause of the increased length of these sessions. In further analysis of the orders placed to the archive, we found that a third of them were for entire broadcasts, while 17% of the orders were for subsections of broadcasts that had been previously defined by archivists. Just under half of the orders were for audiovisual fragments with a start and end time specified by the users themselves. These fragments were typically on the order of a few minutes in duration, with 28% of fragments being one minute or less. In terms of age, orders were for both recent and historical material, with 46% of orders being for items that were broadcast more than one year before the date the order was made.

To answer **CRQ 2**, *what kinds of queries are users issuing to the archive?*, nearly all of the queries contained a keyword search in the form of free text, while almost a quarter specified a date filter. Searching within specific catalog fields was used in 9% of the queries, and frequently specified the media format or copyright owner of the results to be returned. The most frequently occurring keyword searches consisted primarily of program titles. However, these frequently occurring searches accounted for less than 6% of the keyword searches in the archive, leaving a long tail of unaccounted-for queries.

This brings us to **CRQ 3**: *what kinds of terms are contained in the queries issued to the archive?* To address this, we performed a content analysis of the terms in keyword searches. We used catalog information as well as session data to match terms in a keyword search to the titles and thesaurus entries of the documents that were clicked during a user session. Of all the queries where users clicked a result during the session, 41% contained a title term. Thesaurus terms were identified in 44%

of the queries. The faceted thesaurus structure allowed us to further characterize the content of the long tail of unaccounted-for queries. Approximately one quarter of thesaurus terms consisted of general subject keywords such as *soccer*, *election*, and *child*. Another quarter consisted of the names of people, especially the names of politicians and royalty. The remaining terms were classified as locations, program makers, other proper names, or genres. Out of these results we conclude that while titles play a significant role in queries issued to the archive, thesaurus terms describing the content of video also play an important part in searches, with many searches for people and for general subject terms.

3.4.2 Implications for the Thesis

Let us zoom out and consider the broader implications of our findings in the context of this thesis. Qualitative studies have shown that both archive catalogers [165] and professional archive searchers [171] view shot-level annotation of video material as desirable. Meanwhile, we saw that while media professionals frequently order small fragments of broadcasts, the archive search engine returns entire broadcasts or (when available) predefined story segments as results. It takes the searchers longer to order a fragment of a broadcast than it does for them to order a whole broadcast. We conjecture that this time to order a fragment could be cut down by annotating at the shot-level and using them to enable fine-grained search within audiovisual broadcasts. While human annotation of shots is not feasible when considering the number of broadcasts flowing into the archives of today [165], automatic content-based video analysis methods provide a potential solution. However, further research is necessary to evaluate the extent to which these could aid retrieval in the audiovisual broadcast archive. We will present our own contributions in this area in Part II of the thesis.

While the transaction logs of the Netherlands Institute for Sound and Vision provide a rich source of data for study and experimentation, they cannot be shared with the research community at large because of privacy considerations. In the next chapter we will explore a potential alternative to releasing the original query data. Namely, we explore the simulation of queries from the archive. Simulation can provide us with a means to validate retrieval models and ideas about user behavior, as we saw in Chapter 2. In addition, in the context of privacy-sensitive query logs, simulation might provide a way to release artificial — but realistic — sets of queries to the public.

Chapter 4

Simulating Logged User Queries

Transaction logs provide a rich source of data, not only for studying the characteristics of searchers in a real-world environment, but also as material for experiments that aim to take real-world searches and searchers into account [1, 27, 86, 188]. In this chapter we will perform such an experiment. Namely, we investigate the creation of a simulator to produce the queries and orders that are similar to those recorded in the transaction logs of the Netherlands Institute for Sound and Vision.

As we saw in Chapter 2, simulation is a method by which user-based experiments may be performed when real-world information is either unobtainable or too expensive to acquire. This makes simulation a valuable solution for evaluating information retrieval (IR) theories and systems, especially for interactive settings. In the context of transaction logs, an increasing number of studies have been performed into how searches and result clicks recorded in transaction logs may be used to evaluate the performance of retrieval systems [58, 62, 85, 124]. However, transaction logs often contain information that can be used to breach the privacy of search engine users [9]. In addition, their content may contain commercially sensitive information. Therefore organizations are reluctant to release such data for open benchmarking activities.

A solution may lie in using simulation to generate artificial user queries and judgments. In this way large numbers of queries and judgments may be obtained without need for human action, as a substitute for hand-crafted individual retrieval queries and explicitly judged sets of relevant documents. Simulated queries have been compared to manually created queries for information retrieval [8, 162]. Reproducing absolute evaluation scores through simulation has been found to be challenging, as absolute scores will change with, e.g., the recall base. Reproducing exact retrieval scores is not essential to developing a useful simulator when we wish to rank re-

trieval systems by their performance. Our aim is to make a simulator that allows us to identify the *best performing* retrieval system. Consequently, we assess simulation approaches based on how well they predict the relative performance of different retrieval systems. More precisely, we examine whether simulated queries and relevance judgments can be used to create an artificial evaluation testbed that reliably ranks retrieval systems according to their performance.

The work in this chapter is aimed at answering the question,

> **RQ 2** *Can we recreate those searches by media professionals that result in purchases, and use them to create an artificial testbed for retrieval evaluation?*

Here we are directed by the following research questions:

CRQ 1 How well do system rankings obtained using simulated queries correlate with system rankings using logged user queries?

CRQ 2 What factors impact the validity of a simulator?

CRQ 3 Does a simulator that incorporates knowledge about searchers' behavior correlate more closely with rankings based on logged user queries than a simulator that does not?

In the simulation approaches that we will introduce below, we integrate insights into users' search goals and strategies to improve the simulation, e.g., patterns regarding the items typically sought and/or purchased. We also enrich the simulators by incorporating characteristics of the document collection in the commercial environment in which we work, in particular the fielded nature of the documents.

Purchase-query pairs (i.e., user queries and their subsequent orders) can be identified from the transaction logs described in the previous chapter, and we use these purchases as implicit positive relevance judgments for the associated queries [62]. In addition, we use a set of training purchase-query pairs to help inform our simulators. We use the simulators to create sets of artificial purchase-query pairs to use as evaluation testbeds. Each simulator is then assessed by determining whether its testbed produces similar retrieval system rankings to the gold standard testbed created using purchase-query pairs taken from logged user queries.

The main contributions of this chapter are:

- A large-scale study of the correlation between system rankings derived using simulated purchase-query pairs and a system ranking obtained using real queries and implicit assessments.
- Novel query simulation methods that exploit characteristics from real queries as well as document structure and collection structure.

- An analysis of factors that impact the validity of a simulator.

We will use the findings of this chapter in Chapter 10, where we generate a set of artificial purchase-query pairs with our best performing simulator, and use these to evaluate the performance of a system equipped with content-based video retrieval techniques. Furthermore, our framework and findings can be used by creators of evaluation collections who have access to transaction log data but are unable to release it and by experimenters who want to create realistic queries for a document collection without having access to the transaction logs.

We next describe related work on simulation for information retrieval evaluation purposes in Section 4.1. In Section 4.2 we describe our simulation framework and the simulator validation methodology. This is followed by a description of the experimental setup in Section 4.3 and results in Section 4.4. Conclusions for this chapter are presented in Section 4.5.

4.1 Related Work

Our review in Chapter 2 showed that simulation is sometimes used to artificially recreate searcher actions in the field of audiovisual search, on the basis of a predefined set of queries and relevance judgments. Here we provide an overview of related work aimed at simulating the queries and relevance judgments themselves.

Research into simulation for information retrieval systems goes back to at least the 1970s [47]. Early work focused on simulating not only queries and relevance assessments, but also the documents in a collection. Tague et al. [163] developed simulation models that produced output data similar to real-world data, as measured for example by term frequency distributions. However, retrieval experiments showed that evaluations using the simulated data did not score retrieval systems in the same way as evaluations using real-world data [162]. Gordon [46] suggested that simulators should focus on creating queries and relevance judgments for pre-existing (non-simulated) document collections: the remainder of the work that we describe has been performed under this condition.

A problem when simulating queries is to identify multiple relevant documents for a query. One solution has been to use document labels or pre-existing relevance judgments to identify sets of multiple related documents in a collection [7, 65, 87, 93]. Queries are back-generated from the related documents, which are then considered relevant for that query. This allows queries to be defined so that they conform to predefined criteria, e.g., long vs. short. While queries generated in this manner can be used to evaluate retrieval systems, for validation purposes there is typically no user interaction data available. However, it is unclear how they compare to "real-world" queries, and therefore whether the findings based on such queries still hold

when faced with human searchers.

An alternative to identifying multiple relevant documents per query is to focus on identifying a single relevant document per query. Dang and Croft [29] work under this condition in the context of online search. They exploit hyperlinked anchor texts, using the anchor text as a simulated query and the hyperlink target as a relevant document.

Azzopardi et al. [8] study the building of simulators for known-item queries — queries where the user is interested in finding a single specific document. Their approach is to first select a target known-item document from the collection, and to subsequently back-generate a query from the terms in the selected document. In both of these studies, simulator output was compared to real-world data with respect to validity for evaluating retrieval systems. Dang and Croft [29] modified their simulated queries and compared system evaluations on those queries to system evaluations on similarly modified queries and clicks taken from an actual search log. They found that, in terms of system evaluation, the simulated queries reacted similarly to various modification strategies as real queries. Azzopardi et al. [8] compared their simulated known-item queries to sets of 100 manually created known-item queries, examining absolute evaluation scores attained by evaluating systems on the simulator output data. The simulator output sometimes produced retrieval scores statistically indistinguishable to those produced using real known-item queries.

Factors found to affect the simulator assessments include the strategy used to select known items, and the strategy used to select query terms from a target document. We will investigate the effect of altering both strategies in our own experiments.

Our work is similar to that of Azzopardi et al. [8] as our goal is to create valid simulated queries for retrieval evaluation. However, we focus on accurate ranking of retrieval systems, rather than reproducing absolute retrieval evaluation scores for individual retrieval systems. Here we are motivated by experiments with human query creators, which have shown that comparative evaluation of retrieval performance across query creators may be stable even when there are substantial differences in evaluation scores for individual systems [177]. We view each simulator as a query creator (and purchase creator), and the purchase-query pairs contained in transaction logs as having been created by a "gold standard" query creator. The gold standard set of purchase-query pairs is then used to validate simulators. We also differ to Azzopardi et al. [8] in that we base our simulator validation on implicit queries and relevance judgments of audiovisual searchers in transaction logs, rather than explicit queries and judgments generated by a human query creator. We apply and extend simulation strategies developed in [8], and compare these to strategies that take characteristics of the logged queries into account.

4.2 Simulation Framework

First, we describe our setup for assessing query/document pair simulators. Then we detail the simulators, including how they select target documents and query terms.

4.2.1 Validating Simulators

In validating simulators we are particularly interested in how closely the system rankings produced using simulated purchase-query pairs resemble the system rankings produced using real purchase-query pairs. To assess a simulator, we first run it to create an evaluation testbed consisting of a set of queries with associated relevant documents (one relevant document per query, thus resembling a known-item task).

The retrieval systems are then run on the simulated queries in the evaluation testbed, and the associated relevance judgments are used to score the results. We evaluate the performance of each retrieval system on the in terms of Mean Reciprocal Rank (MRR). MRR is a standard evaluation measure for tasks where there is only one relevant item per query [26, 174]. Reciprocal rank for a given query is the inverse of the rank of the first correct result; MRR is the average of the reciprocal ranks for all queries. For the experiments in this chapter, the systems are ranked by their MRR scores.

Once a ranking of systems has been obtained on the simulated evaluation testbed, this ranking is compared to one obtained by ranking the systems using real purchase-query pairs taken from our transaction log. Comparisons between system rankings are couched in terms of the rank correlation coefficient, Kendall's τ. A "better" (read: more realistic) simulator would achieve a higher rank correlation score with the gold standard ranking. Kendall's τ is based on the notion of *concordance* — if a pair of items is ranked in the same order in two lists, it is concordant. If the pair is ranked in opposite order, it is discordant. The rank correlation is given by

$$\tau = \frac{n(Q) - n(D)}{n(Q) + n(D)}, \tag{4.1}$$

where $n(Q)$ is the number of concordant pairs, and $n(D)$ is the number of discordant pairs [112]. When two ranked lists are identical $\tau = 1$, and when one ranked list is the inverse of the other $\tau = 0$. The expected correlation of two rankings chosen at random is 0. Kendall's τ has been used previously for determining the similarity of system rankings on evaluation testbeds generated by different sets of human assessors. Voorhees [177] considered evaluation testbeds with $\tau \geq 0.9$ to be equivalent. In our own setting, we are considering the similarity of system rankings on a testbed that has been created by a simulator to a gold standard testbed that has been created by humans, as τ increases, the rankings produced by the simulated

Algorithm 1 Overview of our purchase-query pair simulator

Initialize an empty query $q = \{\}$
Select the document d_p to be the purchased document with probability $p(d_p)$
Select the query length l with probability $p(l)$
for i in 1 .. l **do**
  if Using field priors **then**
    Select the document field f from which a term should be sampled with probability $p(f)$
    Select a term t_i from the field model (θ_f) of f with probability $p(t_i|\theta_f)$
  else
    Select a term t_i from the document model (θ_d) with probability $p(t_i|\theta_d)$
  Add t_i to the query q
Record d_p and q to define the purchase-query pair

testbed become more similar to the rankings produced by the gold standard testbed, when this occurs we say the simulator becomes more valid.

4.2.2 Simulating Purchase Queries

We base our framework for simulating purchase-query pairs on the known-item search simulation framework presented in [8]. This choice is motivated by the similarity of search for purchase queries to known-item searches, in that for both types of search, a single query is (usually) associated with a single relevant document. For purchase queries the user may not necessarily know beforehand exactly which item he or she wishes to obtain, but will usually purchase a single item. In an analysis of our transaction log, we found that 92% of the purchase queries led to exactly one purchased item.

We extend the framework of Azzopardi et al. [8] by incorporating information about the document fields from which query terms are typically drawn (detailed below). We observed in Chapter 3 that users of search systems tend to select query terms from specific parts of a document. To take this into account, we allow the simulator to use information about document fields to select query terms.

The resulting simulator framework is summarized in Algorithm 1. First, select a document d_p from the collection C that is considered to be the target document to be purchased, at random from a distribution D_d. We determine the length of the query by drawing from a distribution D_l that is estimated using a random subset of the development data (a sample of real purchase queries). For each term to be added to the query, we then determine the field from which the term should be drawn according to distribution D_f, and finally sample a term from the associated term distribution D_t.

In this setting, the crucial problems become: (1) determining D_d, i.e., which document should to be the target of a query; (2) determining D_f, i.e., which field should be selected as the source for a query term; and (3) determining D_t, i.e., which terms should be used to search for the target document. We propose and compare multiple strategies for addressing each problem, which we will discuss in turn.

D_d: Selecting target documents

We investigate the effect of varying the selection of documents to use as simulated purchases. In previous work, selecting target documents according to document importance as captured by inlink counts was found to have a positive effect in obtaining scores closer to retrieval scores using "real" queries [8]. We operate in an environment where inlink information is not available. Therefore, we formulate two target selection strategies that are expected to be representative of a lower and upper bound in simulation accuracy: (1) a uniform selection strategy, and (2) an oracle selection strategy that selects documents that are known to have been purchased.

Uniform All documents in the collection are equally likely to be selected (samples are drawn with replacement). This strategy only requires the presence of a document collection and does not assume additional information. The probability of selecting a document is

$$p_{\text{Uni}}(d_p) = \frac{1}{|C|}, \tag{4.2}$$

where $|C|$ is the collection size.

Oracle For each logged purchased document, a query is generated. This strategy exactly mirrors the distribution of purchased documents that is observed in the test collection. The probability of selecting a document is determined by

$$p_{\text{Ora}}(d_p) = \frac{n(d_p)}{(\sum_{d \in D_p} n(d))}, \tag{4.3}$$

where $n(\cdot)$ is the number of times a document has been purchased and D_p is the set of purchased documents.

We expect the oracle selection strategy to result in a simulator for which the resulting system rankings more closely resemble a ranking resulting from real queries. If the two document selection strategies lead to large differences in correlations with a system ranking produced by real queries, this would mean that more complex strategies for generating purchase distributions should be investigated further.

D_t: Selecting query terms

The second step for a purchase-query simulator is to generate query terms that a user might use to (re)find a given target document. Many strategies are possible — we focus on the effect of existing term selection methods and the effect of selecting terms from different document fields. The following selection methods, previously defined for known-item search in [8], are investigated:

Popular Query terms are sampled from the purchased document using the maximum likelihood estimator. Frequently occurring terms in the document are most likely to be sampled. The probability of sampling a term is determined by:

$$p(t_i|\theta_d) = \frac{n(t_i, d)}{\sum_{t_j \in d} n(t_j, d)}, \tag{4.4}$$

where $n(t, d)$ is the number of times t occurs in d.

Uniform Query terms are sampled from the document using a uniform distribution (each term has an equally likely chance of being sampled):

$$p(t_i|\theta_d) = \frac{1}{|d|}, \tag{4.5}$$

where $|d|$ is the number of unique terms in a document.

Discriminative Query terms are sampled from the document using their inverse collection frequency. Terms that rarely occur in the collection are most likely to be sampled. The probability of sampling these terms is determined by:

$$p(t_i|\theta_d) = \frac{b(t_i, d)}{p(t_i) \cdot \sum_{t_j \in d} \frac{b(t_j, d)}{p(t_j)}}, \tag{4.6}$$

where $b(t, d)$ is an indicator function set to 1 if t is present in d and 0 otherwise, and $p(t)$ is given by:

$$p(t_i) = \frac{\sum_{d \in C} n(t_i, d)}{\sum_{d \in C} \sum_{t_j \in d} n(t_j, d)}. \tag{4.7}$$

TF.IDF Query terms are sampled from the document according to their Term Frequency - Inverse Document Frequency (TF.IDF) value. Terms that occur rarely in the collection, but frequently in the document, are most likely to be sampled. Writing $df(t)$ for the document frequency of a term (i.e., the number of

documents in C containing t, we put:

$$p(t_i|\theta_d) = \frac{n(t_i, d) \cdot \log \frac{|C|}{df(t_i)}}{\sum_{t_j \in d} n(t_j, d) \log \frac{|C|}{df(t_j)}}. \tag{4.8}$$

We restrict ourselves to these four term sampling strategies in order to compare our work to that of [8]. Sampling strategies are expected to affect how realistic simulated queries are, as they constitute different models of how users create query terms when thinking about a target document. Query formulation is more complex in real life, but a model that well explains a large part of real query generation will result in a better simulator.

A simulator that uses a term selection method that is close to that of the term scoring method underlying a retrieval system used to evaluate that simulator will likely score high on the queries thus generated. This does not necessarily result in a good simulator, as we are comparing system rankings to those resulting from evaluation using real purchase queries.

D_f: Incorporating document structure

Beyond the influence of the term selection strategy, we observed in the analysis of query terms in Section 3.3.3 that users of the archive tend to select query terms from specific parts of a document, such as the title. In the collection that we used for development we observed that the program description was the most frequent source of query terms, followed by the summary, title, and recording number fields. Table 4.1 gives a description of the fields that are used in our document collection, and specifies the probability $p(f)$ that query terms are matched in the respective fields. We obtained the probabilities by matching terms in queries from a development set (described in the next section) to their field locations in the associated purchased documents. If a term occurred in multiple fields, each field was counted once. Note that this analysis is not representative of the entire group of queries that was studied in Chapter 3, as only queries that resulted in a purchase have been included in this chapter.

In our simulation experiments, we systematically explore the use of document fields for generating simulated queries. We experiment with restricting the selection of query terms to individual fields on the one hand, and with using the distribution of query terms across fields in our development set on the other hand. When restricting the simulator to individual field, we use the four fields that are the most frequent sources of query terms, as shown in Table 4.1. We also include a simulator that selects terms from the entire document indiscriminately. This gives rise to the following settings:

Field overview			Field overview		
Name	*Description*	$p(f)$	*Name*	*Description*	$p(f)$
beschrijving	Program description	0.448	immix_docid	Document id	0.004
dragernummer	Recording number	0.130	zendgemachtigde	Broadcaster	0.002
hoofdtitel	Program title	0.169	rechten	Copyright holder	0.001
samenvatting	Summary	0.141	genre	Genre	0.001
selectietitel	Story title	0.035	dragertype	Recording format	0.001
makers	Program maker	0.024	deelcatalogus	Sub-collection	0.000
namen	Name	0.019	dragerid	Recording id	0.000
persoonsnamen	Person	0.015	dragersoort	Recording type	0.000
geografische_namen	Location	0.005	herkomst	Origin	0.000
trefwoorden	Subject	0.005	publid	Publication id	0.000

Table 4.1: Names and descriptions of fields available in our experimental document collection; $p(f)$ indicates the probability that the field contains a purchase-query term, derived by matching query terms to terms in purchased documents.

Whole document Query terms are sampled from the entire document without incorporating any information about fields.

Description Query terms are sampled from the description field only.

Summary Query terms are sampled from the summary field only.

Title Query terms are sampled from the title field only.

Recording number Query terms are sampled from the recording number field only.

Field priors Query terms are drawn from any part of the document, but terms from fields that are more frequent sources of real query terms are more likely to be used as a source for query terms. This model corresponds to the full setup outlined in Algorithm 1, including the optional field selection step. Field prior probabilities are estimated using the development collection, and correspond to $p(f)$ in Table 4.1.

4.3 Experimental Setup

We describe the setup of our experiments in terms of the data used, the settings used when applying our simulators, and the retrieval systems used to assess the simulators.

4.3.1 Experimental Data

Our experimental data consists of a collection of documents and a large set of purchase-query pairs. Our documents and purchase-query pairs are obtained from the Netherlands Institute for Sound and Vision. The transaction log data used to identify the purchase-query pairs was gathered between November 18, 2008 and February 1, 2010 (due to the timing of this chapter, the transaction logs used here cover a slightly larger period than the logs analyzed in Chapter 3). In some cases purchases are made for fragments of a broadcast; following the practice at TREC and other evaluation forums [50], we consider the entire broadcast relevant if it contains relevant information, i.e., if a fragment has been purchased. The transaction log includes queries that contain date filters and advanced search terms. We exclude these types of queries from our experiments, leaving their simulation for future work. In some cases a query resulted in purchases for multiple broadcasts, here we consider each purchase-query pair separately. In total we derived 31,237 keyword-only purchase-query pairs from the collection.

Our documents and purchase-query pairs are divided into a development and a test set. The development set is used to derive probability estimates for simulation purposes. The test set is used to produce the gold standard ranking of retrieval systems. In order to preserve the same distribution of documents and purchase-query pairs in the development and test set, documents were randomly assigned to either set. Purchase-query pairs were then assigned to a set depending on the assignments of the purchased document.

4.3.2 Simulator Settings

We create a simulator for each combination of the target, term, and field selection models described in Section 4.2.2. Query lengths for the simulators are drawn from the distribution of query lengths in the development queries. Query terms are taken directly from documents without modification. We generate 1,000 purchase-query pairs per simulator. These are then used to evaluate the retrieval systems described below.

4.3.3 Retrieval Systems

To produce system rankings, we need multiple retrieval systems that generate diverse retrieval runs. To this end we use a variety of indexing and preprocessing methods and scoring functions. We use two open source toolkits for indexing and retrieval: Indri[1] (based on language modeling) and Lucene[2] (based on the vector-

[1] http://www.lemurproject.org/indri/
[2] http://lucene.apache.org/

space model) to create a total of 36 retrieval systems. The main difference between the toolkits is the document scoring method, but they also differ in terms of pre-processing and indexing strategy; both are frequently used in real-world search applications and retrieval experiments. Some of the 36 retrieval systems are designed to give very similar performance scores, to capture subtle differences in ranking of similar systems; others are designed to give very different retrieval performance scores, to capture large fluctuations in ranking.

We build three indexes, one using Indri, and two using Lucene. For the Indri index we use the standard pre-processing settings, without stemming or stop word removal. For the first Lucene index, we apply a standard tokenizer, for the second we additionally remove stop words and apply stemming using the Snowball Stemmer for Dutch.

The Indri retrieval toolkit is based on language modeling and allows for experimentation with different smoothing methods, which we use to generate runs based on a number of models. Documents receive a score that is based on the probability that the language model that generated the query also generated the document. This probability estimate is typically smoothed with term distributions obtained from the collection. The setting of these smoothing parameters can have a large impact on retrieval performance [197]. Here we generate retrieval systems using Dirichlet smoothing with the parameter range $\mu = 50, 250, 500, 1250, 2500, 5000$. In this manner, we generate a total of 6 smoothing-based retrieval runs. These 6 systems can be expected to produce similar retrieval results, allowing us to capture small differences in system rankings.

Both Indri and Lucene provide methods for indexing per field, allowing us to create alternative retrieval systems by forming different combinations of fields; Table 4.1 shows the names and descriptions of the indexed fields. We generate ten fielded retrieval runs for each index (for a total of 30 runs), based on one or more of the following fields: *content* (all text associated with a document), *free(text)* (summary and description), *meta* (title and technical metadata), and *tags* (named entities, genre). The ten field combinations can be expected to give very different performance, while applying a specific field combination to three index types will result in smaller variations in performance.

4.4 Results and Analysis

Our experiments are designed to validate simulation approaches by assessing how well their simulated purchase-query pairs rank retrieval systems in terms of the systems' performance scores. A second goal is to identify the best performing simulator, i.e., the simulator that results in rankings that are closest to the gold standard rank-

	***Uniform* Target Model**				***Oracle* Target Model**			
	Term Model				*Term Model*			
Field Model	*Popular*	*Uniform*	*Discrim.*	*TF.IDF*	*Popular*	*Uniform*	*Discrim.*	*TF.IDF*
Whole Doc.	0.714	0.741	0.666	0.690	0.650	0.697	0.667	0.677
Description	0.393	0.348	0.347	0.360	0.382	0.373	0.371	0.375
Summary	0.352	0.340	0.365	0.355	0.435	0.476	0.286	0.384
Title	0.444	0.461	0.418	0.432	0.385	0.373	0.405	0.392
Recording No.	0.654	0.682	0.645	0.674	0.684	0.704	0.673	0.686
Field Priors	0.738	0.745	0.714	**0.758**	0.738	0.721	0.624	0.687

Table 4.2: Correlation coefficients of system rankings using simulated queries and a system ranking using real-world data. The simulator with the highest coefficient overall is highlighted in bold. Shading has been used to differentiate between ranges of correlation coefficients: darkest shading for $\tau \geq 0.7$, medium shading for $0.5 \leq \tau < 0.7$, light shading for $0.3 \leq \tau < 0.5$, and no shading for $\tau < 0.3$.

ing produced using real queries. In this section we provide an overview and analysis of our experimental results.

The correlation coefficients for the simulators produced in our experiments are given in Table 4.2. The simulator with the lowest coefficient of 0.286 uses *Discriminative* term selection in combination with *Summary* field selection and *Oracle* target selection, indicating that this simulator setting is particularly unsuited for generating realistic purchase-query pairs. The simulator with the highest correlation coefficient of 0.758 uses *Field Prior* field selection in combination with *Uniform* target selection, and *TF.IDF* term selection. None of the simulators achieves the value of $\tau \geq 0.9$ that indicates the equivalence of two testbeds created by human assessors, indicating that there is still plenty of room for improvement in creating simulators that realistically reproduce human querying and purchasing behavior. However, the variance in assessments of the validity of simulator output does give some insight into which simulator strategies are preferable in the framework that we employ.

Incorporating field priors Overall, we can see that simulators incorporating *Field Prior* field selection produce the most reliable system rankings, as measured by correlation to a system ranking using real purchase-query pairs. Except for one case, using field priors consistently and substantially improves over sampling from the whole document without taking field information into account.

The least reliable system rankings are produced by restricting field selection to a single field, as is the case for the *Title*, *Summary*, and *Description* field selection models. An exception is the *Recording Number* field. Simulators using this field selection model achieve correlations that are, depending on the term and target selection models, in many cases similar to the correlations achieved when using the whole document. This points to an important aspect of the real queries, namely, that

many of them are high precision in nature. As we saw in the previous chapter, over 50% of the sessions recorded in the transaction logs of the archive consist of a single query, and a single result click. Professional users often know very well what they are looking for and how to find it, and searching on the highly distinctive recording number allows a document to be quickly targeted. Simulators missing this high-precision field lose out when trying to predict which retrieval systems will perform well in this setting.

Selecting known purchased documents Simulators selecting documents that are known to be purchased (i.e., using *Oracle* target selection) generally do not produce better evaluation testbeds than simulators that select documents uniformly from the entire collection. This is somewhat surprising as Azzopardi et al. [8] found that a non-uniform sampling strategy based on document importance produced better simulators. However, the cited work validated simulators according to their ability to reproduce the absolute retrieval evaluation scores, while in this work we validate simulators according to their ability to rank retrieval systems. This may account for the difference in our findings, although further work is necessary to confirm this.

The impact of term selection Comparing simulator assessment scores across the term selection models, we find that no single model scores high consistently. In our experiments, then, the optimal term selection model is dependent on the target selection and field selection models that are used.

4.5 Conclusion

In this chapter we have studied the design and validation of simulators for generating purchase-query pairs, consisting of a query and an associated relevant purchased document, in a commercial setting. We developed a purchase query simulation framework incorporating new and previously existing simulation approaches. The framework allows us to incorporate different models for: (1) selecting target purchased documents; (2) selecting query terms; and (3) incorporating document structure in the query term selection process. By varying these models we created 48 simulators. Each simulator was used to produce an artificial evaluation testbed of simulated purchase-query pairs. The simulators were validated in terms of how well their testbeds ranked retrieval systems, as compared to the gold standard ranking obtained using a testbed of real logged purchase-query pairs.

For illustration purposes we briefly discuss absolute retrieval scores. Figure 4.1 plots the retrieval scores of retrieval systems evaluated with real purchase-query pairs from transaction logs, against the retrieval scores of the same systems when

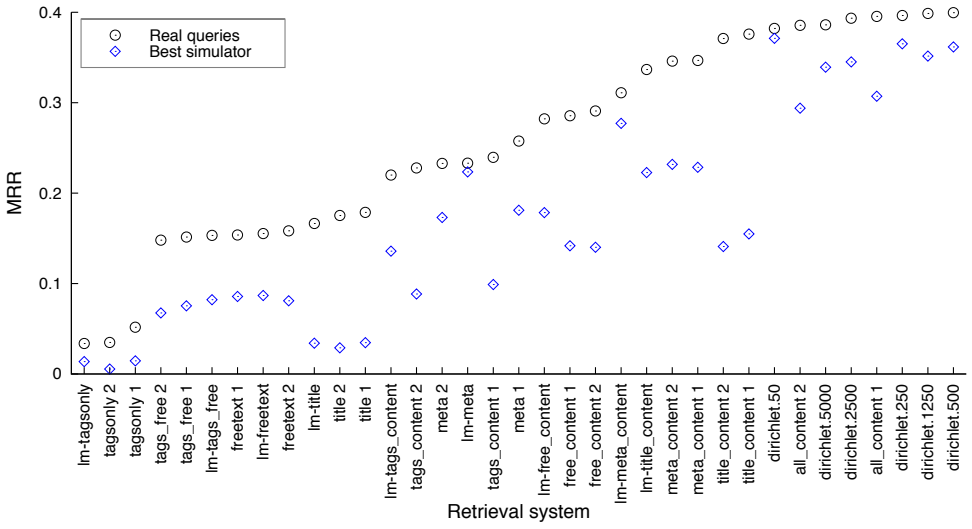


Figure 4.1: MRR scores of 36 retrieval systems on two evaluation testbeds containing purchase-query pairs. *Real queries* plots scores on a testbed taken from the transaction logs of the archive, and *Best simulator* plots scores on a testbed generated by the best simulator ($D_d = Uniform, D_t = Uniform, D_f = FieldPriors$). Retrieval systems are sorted by their performance on real purchase-query pairs.

applied to the artificial purchase-query pairs generated by the best simulator. Some clusters of system rankings are correctly reproduced by the simulator (systems scoring better on real queries also tend to perform better on the simulated queries, and vice versa), even though absolute retrieval scores are generally lower than those obtained using the real queries.

Turning to the research questions that we posed at the beginning of the chapter, with regards to CRQ 1, *How well do system rankings obtained using simulated queries correlate with system rankings using logged user queries?*, no simulator produced an evaluation testbed that, according to the commonly accepted threshold for significance, ranked retrieval systems equivalently to the gold standard purchase-query pairs obtained from transaction logs. This indicates that there is still plenty of room for improvement in the intrinsically difficult query simulation task. However, the variation in assessments of the validity of simulator output did highlight broad categories of more successful and less successful approaches, which help us answer CRQ 2 and CRQ 3. In response to CRQ 2, *What factors impact the validity of a simulator?*, we investigated four models for selecting terms from a purchase document, and found that there was no single term selection model that consistently produced more valid simulators than other term selection models. Further, in our setting, uniform selection of target purchased documents worked at least as well as

a non-uniform selection of known purchased documents. This contrasts with previous findings in a different setting, where selecting "important" documents as targets for known-item searches resulted in improved simulator validity; we leave further investigation of this effect to future work. Finally, to address CRQ 3, *Does a simulator that incorporates knowledge about searchers' behavior correlate more closely with rankings based on logged user queries than a simulator that does not?*, we can reply in the affirmative; simulators were helped by explicitly including information about the distribution of query terms entered by searchers across the different fields of the purchased documents

A previously mentioned advantage of using simulators to create queries and relevance judgments is that, with a simulator, one can generate as many queries as one wishes. In Chapter 10 of this thesis we perform experiments in audiovisual retrieval, and include as one of our collections a set of documents, queries, and relevance judgments taken from the Netherlands Institute for Sound and Vision. Here we employ the best simulator from this chapter as one method for obtaining queries for video retrieval evaluation.

Chapter 5

Conclusion to Part I

Part I of this thesis has focused on deepening our understanding of how professional users search through the audiovisual broadcast archive.

In Chapter 2, we placed our research in context by reviewing previous studies of searchers for video material. Among our findings, we observed that previous studies of searchers in audiovisual archives were small in scale, manually analyzing and categorizing information requests issued to archivists.

In response to this observation, Chapter 3 was devoted to a large-scale study of professional searchers in the archive, by means of a transaction log analysis. From the study it became apparent that the users of the particular audiovisual archive that we studied—the Netherlands Institute for Sound and Vision—demonstrate a high level of expertise. In sessions where an order is made, users often only issue one query and view only one result before obtaining the audiovisual item. This, in combination with the high proportion of searches on program titles, and on specific dates, implies that the users often perform known-item (or target) search, knowing exactly which audiovisual item it is that they wish to obtain when initiating a session. In addition, users order small fragments of broadcasts, but can only retrieve whole broadcasts. Our findings imply that audiovisual archives would be well served by incorporating metadata that allows users to search within individual videos.

In Chapter 4 we investigated whether we could use our new understanding of media professionals to simulate their searches and purchases. We developed a framework for building and validating simulators of searches and purchases, one that incorporates knowledge about the fields of documents in which query terms occur, We used this to framework to develop multiple simulators. We then validated the output of each simulator by determining how closely the rankings of retrieval systems on simulated queries and purchases correlated to the rankings of retrieval systems on

real queries and purchases. The best simulator incorporated knowledge about the distribution of terms in searcher queries across the fields of the documents that they purchased, and achieved a correlation of 0.758. While the correlation with rankings using real queries is high, we believe there is still room for improvement, and feel there is much potential for further research in this area.

Part II
Content-Based Video Retrieval for Audiovisual Broadcast Archives

Progress in digital recording, storage, and networking technology has enabled large-scale ingestion and dissemination of multimedia material. As a consequence, the audiovisual archives responsible for safeguarding the cultural heritage captured in broadcast television recordings are growing rapidly. Traditionally, these archives manually annotate video programs with textual descriptions for preservation and retrieval purposes [38]. Users of the archive search on these textual descriptions and receive (rankings of) complete television programs as results. Fine-grained manual annotation of video fragments is prohibitive, as the work involved is inevitably tedious, incomplete, and costly. Content-based video retrieval may provide a solution. Though imperfect, it offers an abundant source of automatically generated shot-level descriptions for search. Not surprisingly, there is growing interest from audiovisual archives in using content-based video retrieval to supplement their current practice [15].

In Part II of the thesis we will examine in detail how content-based video retrieval methods may be used to help improve search for video. We use the insights that we have gained in Part I of the thesis to place our experiments in the context of the audiovisual broadcast archive. Specifically, we develop a test collection based directly on the searches and purchases that we studied in Chapter 3, and we use a simulator from Chapter 4 to produce additional queries for retrieval evaluation. Contributions

of this part of the thesis include: a detailed examination of the concept selection task for detector-based search; a retrieval model that incorporates redundancy between the mention of query words in video and their appearance on screen; and a study of the potential of various combinations of automatically generated content metadata and manually created archive text to improve retrieval in the audiovisual broadcast archive.

Chapter 6

Methods for Content-Based Video Retrieval

Machine-driven analysis of audiovisual material allows us to automatically create metadata that describes different aspects of within-video content, a process we refer to as *multimedia content analysis* [36, 180]. These machine-generated descriptions can then be used to help searchers locate fragments of video in a collection, in a process encompassed by terms such as content-based video search [18], content-based multimedia information retrieval [99, 183], and *content-based video retrieval* [142, 150]. The latter is the term that we will continue to use in this thesis. In this chapter we will review related work on content-based video retrieval, and, as the some background on the multimedia content analysis methods on which the descriptions are based.

The literature on content-based video retrieval is vast, and it is not our aim to provide an exhaustive overview of the field. Instead, we provide the setting needed for the remainder of the thesis and to be able to assess the contributions that we make. We will start by sketching the benchmarking environment in which much of the recent research in content-based video retrieval has been taking place, and follow this with a summary of the ways in which automatically generated data may be used for search. We then zoom in on two so-called mismatch problems in content-based video retrieval, problems that we will address in Chapter 8 and Chapter 9 of this thesis, before wrapping up with conclusions for this chapter.

6.1 The Role of Benchmarking

Benchmarking activities play a key role in understanding the merits of methods for content-based video retrieval, and in addition, benchmarking methodologies play an important part in the study of content-based video retrieval for the audiovisual archive with which we wrap up the thesis in Chapter 10. Therefore we will describe the role of benchmarking in information retrieval in general, and in content-based video retrieval in particular, before moving on to describe how automatically generated metadata is used for search.

The field of information retrieval is driven by the quantitative evaluation of systems [128]. Benchmarking is a key to developing a fair basis for such evaluation. This is done by developing *test collections* that can be used to assess the performance of individual systems on the same data set. In information retrieval, a test collection is usually understood to comprise of, at minimum, a collection of documents, a set of queries, and a set of relevance judgments indicating the relevance of documents from the collection to each query [134]. A researcher who has access to a benchmark collection can load documents into their retrieval system and then process the query set, and create a ranked list of document results for every query. An evaluation measure is used to assess the performance of a system over all queries. An important aspect of a benchmark collection is that it can be accessed by all members of a research community, thereby serving as a yardstick for comparing different systems. In this way benchmarking serves as a driving force for researchers in a given area.

Progress in content-based video retrieval has been assessed by benchmarking activities that have created video-based test collections that allow for fair comparison of systems. Notable examples of such benchmarking activities include: ETISEO [114] and PETS [41], which focus on surveillance video; the AMI project [125], which emphasizes the analysis of meeting videos; VideoCLEF [98] which studies video retrieval tasks within a multilingual environment; and the US Government Video Analysis and Content Extraction program, which ultimately aims for general video content understanding based on visual recognition, see [89] for a representative example. However, the benchmarking activity that is the most closely related to this thesis is the annual TREC Video Retrieval Evaluation [146] (TRECVID), organized by the American National Institute of Standards and Technology (NIST). Over the years, TRECVID has emphasized retrieval in collections of broadcast news, and, more recently, in collection of audiovisual archive footage. Not only are the test collections produced by TRECVID closely related to the domain of the audiovisual archive, their collections have played a key role in fostering content-based video retrieval progress [39, 153, 191].

TRECVID TRECVID started as a part of TREC, a benchmarking activity that organizes test collections for a diverse range of information retrieval tasks [175], in 2001. It became an independent benchmarking activity in 2003 [145]. The aim of TRECVID, as given in the activity's annual retrieval evaluation guidelines, is "[...] to promote progress in content-based analysis of and retrieval from digital video via open, metrics-based evaluation."[1] To this end it annually defines benchmark collections that are released to participating members. In addition, these collections have been designed so that they can be used to evaluate a variety of tasks, not only content-based video retrieval tasks, but also multimedia content analysis tasks such as concept detection, which we will examine in detail further, and (in the early years) shot segmentation. In the period 2003–2009, the TRECVID test collections have been sourced primarily from two domains: broadcast news (in the collections released from 2003–2006), and the audiovisual broadcast archive (specifically, from the Netherlands Institute for Sound and vision, in the collections released from 2007–2009). Over the years, the TRECVID test collections have been created such that they can be used to evaluate multiple tasks, including shot segmentation, semantic concept detection, copy detection, and surveillance event detection. A recurring task since 2003 has been the *search task*; as this thesis is concerned with the quantitative study of search of video content, we will describe the collections and evaluation setup for the search task in further detail. The test collections for our experiments will be based on collections derived from TRECVID.

Ingredients of a collection As we mentioned earlier, a test collection for information retrieval is commonly understood to consist of a collection of documents, a set of queries, and a set of relevance judgments identifying the relevance of documents for each query. This holds true for the TRECVID test collections, but here the collection of documents is a set of *video documents*, consisting of video data and metadata. The set of queries, in turn, is a set of *multimedia queries*, consisting of both text descriptions and example images and/or videos, which have been created by the benchmark evaluators specifically for content-based video retrieval task (partly on the basis of their perceived difficulty, for example by dividing queries between specific and generic queries). The set of relevance judgments is produced after the video data and queries have been released; participants in the official benchmark evaluation are asked to send results from their retrieval systems to the benchmark organizer, the results are then pooled, and the top results in the pool for each query are viewed by paid judges, who record whether individual shots are relevant or not for a query [94, 116, 117, 144].

In addition to test collections, TRECVID also releases *development* collections.

[1] `http://www-nlpir.nist.gov/projects/tv2010/tv2010.html`

These are designed so that they can be used to learn concept detectors, and are manually labelled by group effort with respect to concept vocabularies [146]. The shared collections have also been enhanced over the years by community donated efforts, e.g., [84, 110, 156]. In the next chapter we will define two collections for our experiments, created on the basis of the TRECVID basis collection and augmented with data from other efforts.

Search tasks Retrieval systems are compared on the basis of the ranked results that are returned in response to the queries that are in the test collection. The retrieval systems may or may not return results with the benefit of human input; TRECVID has defined three search tasks specifying varying levels of human interaction:

Fully automatic The system takes a query as input and produces results without human intervention.

Manually-assisted A human searcher may reformulate the original query *once*. The reformulated query is then passed to the system which produces results without any further intervention.

Interactive A human searcher may reformulate the original query and pass it to the system multiple times, selecting and ranking the final results to be evaluated.

We pointed out in Section 2.3 that it is difficult to assess the merits of retrieval algorithms when there is a human in the loop, as different users on the same system produce results of varying quality according to their expertise. Therefore we will perform all our retrieval experiments under the fully automatic search task setting. In this way the impact of different retrieval strategies can be assessed without bias resulting from user variance.

Evaluation measure An essential component of any retrieval task is evaluating the retrieval performance of a set of results. The two fundamental measures of retrieval performance are *precision* and *recall* [102]. Precision is the fraction of retrieved documents that have been judged to be relevant, and recall is the fraction of all relevant documents that have been retrieved. These two measures are defined for the simple case where a search system returns a set of documents for a query.

Precision and recall do not of themselves take ranking information into account, and therefore do not reward the placement of (many) relevant documents at the top of a result list. Multiple rank-based measures have been proposed [13], one of the most frequently used measures is *average precision* (AP), which combines precision and recall. It is defined as “the mean of the precision scores obtained after

each relevant document is retrieved, using zero as the precision for relevant documents that are not retrieved" [13]. To increase the stability of measurement, AP may be averaged across multiple queries to give *mean average precision* MAP provides combines both precision and recall, rewarding high ranking of relevant results, and has become the de-facto measure for evaluating the ranked result lists produced by content-based video retrieval systems. Accordingly, we will use MAP to evaluate retrieval experiments in this thesis.

6.2 Searching with Automatically Generated Content Metadata

Let us turn now from the benchmarking environment to a discussion of the practice of content-based video retrieval itself. We will organize our discussion along three sources of automatically generated content metadata that have played an important role in enabling search:

- automatically generated *transcripts* of spoken dialogue,
- *low-level features* extracted from the video and audio signals, and,
- *detected concepts* that are estimated on the basis of low-level features using learned concept models.

Another possible source of automatically generated content metadata includes recognized objects within video. These differ to detected concepts in that objects are localized within video, rather than defined by a global estimation. We do not include recognized objects in our review as they were not available to us. However, in any content-based video retrieval systems more than one source of metadata may be available, and in this case multiple result lists will need to be combined to produce a final response. Therefore we will finish this section with a discussion of combination approaches for providing a final list of content-based retrieval results.

Let us note once again that there is an extensive body of work on content-based video retrieval, and we only sketch principal aspects here. For comprehensive overviews we refer the reader to Yan and Hauptmann [191] on transcript-based and feature-based search, Datta et al. [30] for feature-based search, Snoek and Worring [153] for a review of recent developments in detector based search, and Wilkins [183] for a study on the effectiveness of combination approaches for combining result lists.

6.2.1 Transcript-Based Search

Automatic speech recognition is a technology that allows us to transcribe spoken words from an audio signal into textual form. This technology can attain high accuracy rates, with word accuracy rates approaching those of manually created transcripts for high-quality recorded audio in well-defined domains with high quality audio recording such as broadcast news [31]. Accuracy is not as high for less well-defined domains and languages, where word error rates of 50–60% are common [43]. However, even when automatic speech recognition transcripts have high error rates, they can still be effectively applied to the retrieval task [45, 55]. We term the use of the text that is produced by automatic speech recognition technology to search for video *transcript-based search*.

Transcripts consist of text, and therefore text information retrieval methods are commonly used when indexing, querying, and retrieving shots in transcript-based search [191]. There is a characteristic specific to video data that needs to be taken into account; information being presented in the audio signal may not necessarily be aligned with the information in the video signal. For example, while watching a news broadcast, the anchorperson may be introducing a story about an armed conflict taking place in the Middle East, while the video is displaying a shot of the studio. This is one of the two matching problems that we will review in further detail in Section 6.3, the *temporal mismatch*. For now, we will keep to a summarization of the basic steps for indexing, querying, and retrieval in transcript-based search.

When indexing with transcripts, the words from the transcripts are preprocessed and assigned to individual shots in the collection. In the preprocessing step, it is standard practice to normalize words by reducing all letters to lower case and removing punctuation. In addition, frequently occurring stop words (such as “the”, “and”, and “for”) are removed [54, 131, 149, 191]. In order to improve recall, morphological variants of words may be reduced to their canonical forms in a process called *stemming* [54, 149, 191]. In this way the words *running* and *runs*, for example, can be reduced to the root form *run*. The processed words, or terms, are then placed in an index ready for retrieval.

At query time, a text query is passed to the system. The query text is preprocessed in the same way as the transcript text, so that they can be matched against the index. To increase either recall or precision, text-based query expansion may be performed to add, delete, or reweigh words in the query, for example by using a parallel set of text documents to identify words that are closely related to the query [23], by identifying related words in a structured thesaurus [131]. Once the query terms have been preprocessed, we are ready for retrieval.

At retrieval time, the words from the query are matched to words in the shots using a retrieval model. There is a large range of well-established retrieval mod-

els available to choose from in the literature, and a variety of them have been used successfully for transcript-based search. The standard retrieval model used in transcript based search seems to be the OKAPI formula, in particular the BM25 formula [3, 53, 149], (in Chapter 8 we will use BM25 for retrieval with textual labels) but other approaches such as language modeling have also been used with success to retrieve shots on the basis of transcripts [161, 182].

Transcript-based search has met with considerable effectiveness for some queries in content-based video retrieval [94], however, it can only be successful when the narrative of the video provides a spoken commentary on the visual content of the video [143]. In addition the searcher's query must be such that it is likely to be addressed in the video commentary. For example, when searching through broadcast news, transcript-based search is more likely to be more effective for a query such as *Find shots of Hu Jintao, president of China* than for a query such as *Find shots with hills or mountains visible*, as Hu Jintao is a person who is likely to be mentioned in transcripts, while hills and mountains are more likely to be contained only in the video signal. To address situations where transcript-based search is insufficient, we will need to search on other sources of automatically generated content metadata.

6.2.2 Low-Level Feature-Based Search

Low-level features such as color and texture are extracted directly from video data, and serve as a foundation for many content-based video retrieval systems. These features can be used, not only for search, but also as building blocks for developing high-level machine understanding of the semantic content of video. However, in and of themselves they do not impart high-level understanding of video content, hence we refer to them as low-level features. We term search on these feature annotations, *feature-based search*.

In general low-level features can be computed from the video signal (visual features) and the audio signal (audio features). Visual features are computed from either individual keyframes, or, when taking temporal aspects of video into account, calculated across sequences of keyframes extracted from video. Notable types of visual features are *color*, *texture*, and *shape*. These can be further subdivided according to the spatial scale at which they are computed; globally, regionally, at the keypoint level, and across the temporal dimension [153]. Audio features, less often used than visual features in video retrieval, are calculated by their change in amplitude across time, and characterize qualities such as volume, pitch, amplitude, and spectral features [180]. Once computed the features are stored, ready for retrieval.

So how does one formulate a query for a feature-based search system? As low-level features are difficult for humans to interpret, this has represented a challenge for designers of systems that incorporate feature-based search. A number of in-

novative query input methods have been proposed over the years, including structured query languages that allow users to enter desired low-level feature characteristics as textual input ("Find videos with a red, round object") [42], and visual interfaces for specifying colors, textures, objects, and motion that the user wishes to see in the results returned by the system [18]. However, the dominant trend over the years has been to pose queries for feature-based search in the form of *multimedia examples* — images or videos — so that low-level features computed from the examples can be matched against those computed for the shots in the collection [37, 105, 115, 191, 198]. In some cases the searcher may manually specify a subpart of an keyframe in order to further denote the desired object of a search [141].

Once low-level features have been computed and a query has been issued to a content-based video retrieval system, a system needs to return relevant results. Commonly this is done by using distance metrics such as the Euclidean distance measure or histogram intersection to measure the similarity between query features and the features of shots in the collection [191]. In addition, pseudo-relevance feedback and learning approaches can be used to learn the appropriate feature combination adaptively [161, 192]. When multiple results are produced (for example with multiple input query examples, or when considering different combinations of features), feature-based search can produce multiple result lists. These are combined to produce a final result list, see McDonald and Smeaton [104] and Yan and Hauptmann [191] for comparisons of different combination functions for combining feature-based search results.

We will use feature-based search in Chapter 10. As feature-based search is not one of the areas where we intend to improve on in this thesis, we use a state-of-the-art feature-based search implementation donated by the MediaMill group, which incorporates a learning-based approach in combination with random sampling of negative examples to return a single ranked list for feature-based search [159].

6.2.3 Detector-Based Search

An alternative to using low-level features directly for search is to learn configurations of features that describe high-level semantic concepts, in a process called *concept detection.* In general, in concept detection, a concept detector takes a video segment as input, and provides as output an estimation of the probability that a given concept such as *Sky*, *Outdoor*, or *Person* is present in that video segment. In the context of content-based video retrieval, the annotations created by concepts detectors can be used to help determine the relevance of shots to a searcher's query. We term the use of the output of concept detectors for retrieval *detector-based search*.

The life of a concept detector consists of two phases, *training* and *testing* [153]. For the training phase, concept detectors require a set of video data, and manually

created annotations for that data indicating the presence or absence of video in the shots. This labelled collection is used to learn an optimal configuration of low-level features (as described in the previous section) that characterize the appearance of a given concept in the video data. In the testing phase, a new set of video data is passed to the concept detector, which on the basis of the learned optimal configuration estimates the probability that the given concept occurs in the shots in the video data. The output of a detector, then, is a score assigned for each shot for a given concepts.

Current approaches focus on the development of generic concept detection algorithms that can be applied to a range of concepts, e.g., [84, 155]. A challenge that accompanies this approach is deciding which concepts to detect. Manual labeling of video for development is resource-intensive, and therefore the number of concepts for which training data can be provided is limited. An important initiative in this area is the LSCOM effort, which defined a vocabulary of more than 1,000 concepts for describing broadcast news, of which over 400 were annotated on the TRECVID 2005 development collection [110]. In addition, Snoek et al. [156] released annotations for a lexicon of 101 concepts on the same set. This has allowed researchers to investigate the creation of hundreds of detectors.

At retrieval time, then, a query must be translated to the lexicon of concepts for which detectors are available. If the lexicon were sufficiently large, a textual query could be directly matched to descriptions of the available set of detectors (Hauptmann et al. [57] estimates 5,000 concept detectors are necessary for internet-quality search). However, the number of concept detectors is limited. The most successful systems are those in which an expert searcher manually selects the most appropriate concepts from the vocabulary for a query [146]. However, such systems demand a high user investment in the retrieval task, especially when concept vocabularies include hundreds of concepts and there is recognition that video retrieval systems should be easier to use to be successful [143]. Therefore *automatic concept selection*, which does not require user interaction, has received much attention [153]. This brings us to our other mismatch problem, the *vocabulary mismatch* between a user's information need and the available detectors in the concept vocabulary. This is the second of the two mismatch problems that we will review in further detail in Section 6.3. For now, suffice to say that the final query, whether created by an expert or through automatic concept selection, consists of a set of of concept detectors, which may or may not be weighted with respect to their importance for the query.

Given a query consisting of a set of (weighted) selected concepts, and a collection of shots that has been automatically annotated with concept detectors, how do we find relevant shots? In effect each concept detector may be considered a query in itself, i.e. the detector *sky* ranks all shots in a collection with respect to the estimated probability that they contain sky. In the simplest case a query is matched to only

one concept, so that the detector becomes the query, and thereby its probability estimates may be used to return a list of ranked results for that concept [158]. However, this ignores valuable retrieval information that may be present in other detectors. When a query consists of more than one detector, the results from each detector need to be combined. This can be done in combination with results from searches on other sources of automatically generated annotations, or for results based on detector annotations alone. The methodology used in either case is that same, so we will include the discussion of concept detectors in the next section.

In Chapter 7 we will define a collection containing archive footage that has been annotated with scores from 57 concept detectors, which were developed by the MediaMill group [159, 169].

6.2.4 Combining Results from Multiple Methods

Once retrieval has been performed using the above methods, multiple result lists are produced. These result lists need to be combined to produce a single ranked result list. The fusion of results from searches on automatically generated metadata is an ongoing problem, in particular because of the variable retrieval effectiveness of different result lists [146]. To simplify the fusion problem, results are commonly combined *hierarchically*, first at the stage of "retrieval experts" (which generally correspond to the different source-based families of search that we have described above) to produce lists of intermediate results, and then these lists of intermediate results are again combined [92, 104, 189]. Another strategy is to combine all result lists simultaneously; while this approach is less common in the literature due to the increased complexity of choosing weighting combinations, Wilkins et al. [185] have demonstrated that the potential retrieval performance that can be gained by combining result lists is doubled in this way.

Whether fusing hierarchically or simultaneously, a combination function is required to merge the result lists. This may also be accompanied by a score normalization function, as different search methods produce different score distributions that do not necessarily mix well. Wilkins [183] performed an extensive quantitative analysis of different score normalization measures and linear combination operators as applied to video retrieval. In their study, consistently the best results were obtained by using a two-step combination function. In the first step, the scores of results in each result list are normalized so that only ranking information is preserved, using a variation on the Borda normalization [6] method that takes the varying lengths of different result lists into account. Then result lists are then combined using the CombMNZ combination operator [44], which heavily weights documents that occur in more than one result list, with an additional parameter that controls the influence of different input result lists on the final result list.

A problem when performing weighted fusion is determining the optimal weight to assign to each result list. Content-based video retrieval is especially sensitive to the choice of weights, due to the variability of retrieval performance of individual result lists. A popular approach in early work was to use query-independent weighting, e.g., to assign the same weights to each source of retrieval results independently of the query [3, 25, 77, 104]. However, as we observed previously, the effectiveness of a source of retrieval results is dependent on the query. Therefore *query-class dependent* approaches have emerged. Here a set of query classes, such as *Sport* [22] or *Named Person* [193], are defined such that queries falling into a class are expected to have similar responsiveness to the available sources of results. Manual definition of query classes has the problem of requiring expert domain knowledge, and therefore methods have been proposed that create query classes automatically [92, 190]. Optimal combination weights for each class are then learned on a training collection. At retrieval time, incoming queries are automatically analyzed and assigned to a query class, and their result lists are combined using the learned weights.

An alternative to simultaneous (hierarchical) result combination is to use sequential combination methods, which iteratively update retrieval results in a process called reranking [67, 91, 111, 111, 166, 192, 196]. In general, such combination algorithms begin with an initial list of ranked results which is then used as the basis for further combination operations, in a process termed reranking. The top ranked results of the initial result list are used as positive examples for obtaining further ranking information [67, 91, 111, 166, 196]. In addition, negative examples may be selected from the bottom of the result list [192] or at random from the collection [111]. Such methods work when the initial list used for reranking is of reasonable quality.

In Chapter 10 we will perform retrieval with all three content-based video retrieval methods we previously described in this section. As we are concerned with evaluating the potential of content-based video retrieval methods for our specific context, and not with the weighting methodology, we follow Wilkins [183] and use oracle result fusion, optimized on a per query basis on the test set. We use the recommended BordaMax score normalization strategy in combination with weighted CombMNZ fusion to obtain a final result list.

6.3 Mismatches in Content-Based Video Retrieval

There are many mismatches in information retrieval, indeed, search in general can be viewed as trying to correct a mismatch between what a user wants and what a user has. Here we will zoom in on two mismatch problems that we have identified for further examination in Chapter 8 and 9 of the thesis.

6.3.1 The Vocabulary Mismatch when Selecting Concepts

Automatic concept selection We previously mentioned that, in detector-based search, there is mismatch problem when matching a user's information need to the available concept vocabulary. We term the process of automatic translation of query input to the concept vocabulary *automatic concept selection*. The output of automatic concept selection commonly includes both the identifiers for the concepts in the lexicon, and confidence scores indicating how likely they are to be appropriate for a query. Automatic concept selection approaches can be divided into three distinct type: *text-based*, *ontology-based*, and *visual-based* [153].

In text-based selection, a text query is matched to textual descriptions of concepts in the concept vocabulary using standard text retrieval techniques [19, 52, 111, 158, 179]. The textual descriptions of the concepts can vary from one or two sentences created by the vocabulary designers [51, 158, 179], to sets of synonyms [113], to large external corpora [111, 113]. The result is a scored, ranked list of concepts whose descriptions match the query text.

In ontology-based selection, concepts in the concept vocabulary are linked to a structured knowledge repository (typically the lexical database WordNet) [107]. At query time, the text of a query is likewise linked to the repository, and structural relationships are used to determine the similarity of each concept in the vocabulary to the query, for example by using semantic similarity measures such as Resnik's measure of information content or Lesk's similarity measure [51, 52, 111, 113, 158, 181]. To avoid returning all concepts for a query (which introduces noise), the ranked list of concepts may be truncated by using a threshold similarity value [111] or by truncating the list to the top n concepts [51, 113].

Visual-based selection differs to text-based and ontology-based selection in that it relies on multimedia examples to determine which concepts should be selected for a query. The general approach here is to detect concepts in the query examples, and use the detector scores to create a ranked list of selected concepts [100, 111, 179]. These scores may be further modified to give frequently occurring concepts such as *Person* and *Sky* a lower value, as they are expected to be less discriminative for finding relevant shots than less frequently occurring concepts [100].

Ideal concept selection Because of the importance of automatic concept selection in the retrieval process, it is desirable to assess the effectiveness of concept selection methods independently of the final retrieval performance of the content-based video retrieval system. Hauff et al. [52] suggested the use of a set of "ideal" query-to-concept mappings to evaluate automatic concept selection. The automatically selected concepts for a query can then be compared to the ideal query-to-concept mapping to assess whether the appropriate concepts have been selected.

So how can we create mappings between a query on the one hand, and a vocabulary of concepts on the other hand? We identify two approaches in the literature, the first based on *human knowledge*, and the second on *collection knowledge*.

Knowledge about the world is implicit in human minds, and retrieval systems can exploit this knowledge by asking humans to select appropriate concepts for individual queries. Christel and Hauptmann [21] analysed such associations, with concept usefulness to a query rated on a scale of 1–5. Two collections were used, one containing 23 queries and 10 concepts, the other containing 24 queries and 17 concepts. Their work showed inter-annotator agreement to be low, with less than 15% of the mappings being agreed upon by all participants. Neo et al. [113] found agreement to be similarly low in an experiment with 8 queries and 24 concepts. In a later experiment with 24 queries and 101 concepts Hauff et al. [52] also report agreement to be "surprisingly low". To compare, inter-annotator agreement for users assigning concept labels to video fragments can be higher than 95% [173]. In all of the work that we have reviewed, study participants performed the query-to-concept mapping task in isolation, which may be one explanation for the variability in the concept mapping performance. In Chapter 8, we will use focus groups to address the query to-concept mapping problem, so that greater agreement can be reached between experiments through group discussion between participants.

Turning to the use of collections to create query-to-concept mappings, recall that development collections for building concept detectors are manually labelled with respect to the presence or absence of concepts. When such a collection is also labelled with respect to the relevance and non-relevance of queries, generative processes can be used to identify concepts that are strongly associated with a query in the collection. Christel and Hauptmann [21], Lin and Hauptmann [101] and Hauptmann et al. [57] use such collections to identify which concepts are the most strongly associated with a query. The resulting query-to-concept mappings are then used to perform retrieval experiments. Relevant concepts are selected by analysing which shots are relevant to a query, and in turn which concepts are associated with relevant shots. This can be done by, e.g., using mutual information to determine the utility of a concept [101]: a concept is mapped to a query if it reduces uncertainty about a particular shot being relevant to that query. Another approach is to compute the probability a shot is relevant to a query given that the shot is associated with a particular concept, normalising for prior probability that any shot is relevant to the query [21]. The identified concepts are then used to perform retrieval experiments. In our study of the query-to-concept mapping problem in Chapter 9, we will adopt the approach of Hauptmann et al. [57] to create a set of collection-based query-to-concept mappings, and compare these to the mappings created through focus group experiments.

6.3.2 The Temporal Mismatch Between the Audio and Video Signals

We turn now from a mismatch of vocabularies to a mismatch in time. In a collection of written documents, each document can be viewed independently of the others. This is not the case for a video collection that has been separated into shots. Shots on their own do not form an independent story; they contain information, but have been edited so that they form a logical story unit only when they are viewed in sequence [152]. Take, for example, the case of a news broadcast, where a newscaster is shown in a studio introducing a story, and shots depicting the subject of the story are shown after this. This results in a temporal mismatch between the text and the video, which for retrieval purposed needs to be taken into account when associating transcript words to individual shots across time.

A standard approach to compensate for the temporal mismatch is to associate a given word at a given time to shots surrounding that time point, within a neighbourhood of n shots [191]. While this approach improves retrieval performance, it is heuristically driven, and does not take any specific properties of the temporal mismatch into account. In a different approach, for the domain of broadcast news, where automatic methods can be used to identify the boundaries of individual stories, transcript words are associated with all shots within a given story in which they occur. This method has been shown to improve retrieval performance, but it can only be applied in genres where cohesive story units occur. In environments such as audiovisual broadcast archives, which contain programs from a variety of genres such as documentaries, educational programs, soaps, and talk shows, as well as news broadcasts, story units are much less clear cut.

An alternative is suggested by Yang et al. [195], who work in the domain of *person finding*, and are concerned with the task of locating people in broadcast news. They performed an investigation into the time at which the names of twenty people in were recorded in transcripts, and contrasted this with the appearance of these people in the video signal, within a neighbourhood of four shots. They found that on average, in their collection, a person appeared about two seconds after their name is mentioned in the speech of the video. The distribution of each named entity was approximated by a Gaussian model, and the authors found that distributions of people occurring frequently in the news shared similar mean values and standard deviations. They went on to use the distribution models to propagate shot relevance scores, and found that by propagating scores to adjacent shots according to a time-based distribution they were able to increase MAP by 30% over a flat window-based approach. In Chapter 9 we are inspired by Yang et al. [195], and perform a large scale analysis of the temporal mismatch between the occurrence of visual items in transcripts and their occurrence in the video signal, on a collection of broadcast news

and on a collection of archive footage, and within a range of 50 shots on either side of the query. In addition to quantifying the temporal mismatch, we also quantify redundancy within the video signal itself, i.e., the property that the same object is likely to occur in multiple shots in close temporal proximity to each other.

6.4 Conclusions

In this chapter we have provided a background for our experiments studying content-based video retrieval for audiovisual archives. We have sketched the benchmarking environment in which our research takes place, identified three search methods according to the type of automatically generated metadata that they operate on, and discussed the combination of results from multiple sources of automatically generated content metadata for improved retrieval performance. We finished by zooming in on two matching problems in the search methods that we identified; (1) matching queries to lexicons of concepts, and (2) matching transcripts to visually relevant results.

The definition of an appropriate test collection is key for studying retrieval-related tasks. In the next chapter we will define two collections as the basis for our experiments in the rest of the thesis. The collections originate from TRECVID benchmarking activities, and are augmented with a multimedia thesaurus combining two pre-existing concept lexicons, queries that are based on the searches recorded in the transaction logs from the Netherlands Institute of Sound and Vision, and manually created archive text obtained from the same institute. Our experiments in Chapter 9 will examine the problem of matching queries to a concept vocabulary that was expanded on in Section 6.3.1, through an experiment in creating benchmarks of ideal query-to-concept mappings. We then move on to the temporal mismatch problem identified in Section 6.3.2, characterizing the redundancy phenomena that occur in different collections, and develop a retrieval model for transcript-based search that takes these phenomena into account. Finally, in Chapter 10, we implement a content-based video retrieval system and apply it to queries from the audiovisual archive. This system will incorporate detector-based search (using the best automatic concept selection method identified in Chapter 8 for detector-based search), transcript-based search (using the temporally adjusted transcript search model developed in Chapter 9), and feature-based search. In addition it will include search on manual catalog data. These are then contrasted and combined, so that we may evaluate the potential impact of content-based video retrieval methods in the audiovisual archive.

Chapter 7

Two Collections

In this chapter we describe two video collections that will be used to perform content-based experiments in Chapters 8, 9, and 10. The first collection is based on a set of news broadcasts, while the second collection is based on archive footage in the form of broadcasts stored by the Netherlands Institute for Sound and Vision. Each collection consists of video data that has been automatically segmented into shots, a set of queries[1] and relevance judgments, and automatically generated metadata for the shots. In addition, the broadcast news collection has been manually labelled with respect to hundreds of visual concepts, from which we create a multimedia thesaurus of 450 semantically enriched concepts. We have augmented the archive footage collection with queries and relevance judgments obtained from logged searcher actions. An overview of the two collections is given in Table 7.1; we will describe each in more detail below.

7.1 The Broadcast News Collection

The transaction log study in Chapter 3 revealed that news bulletin programs — often called "journaal" in Dutch — are a significant component of searches in the audiovisual broadcast archive that we studied. Indeed, the word "journaal" accounted for 12.5% of all title query terms as summarized in Table 3.6. News, then, is an important source of material for professionals searching in the audiovisual broadcast archive.

Following on from this, our first video collection, the *Broadcast News* collection, is based on a set of news bulletin video programs. The shots in this data set have been

[1] Descriptions of the queries contained in the two collections can be found in Appendix C.

Table 7.1: Ingredients of the two video collections used as a basis for the experiments outlined in Chapters 8, 9, and 10

	Video Collection	
Ingredient	*Broadcast News*	*Archive Footage*
Video data	70 hours English, Chinese, and Arabic news broadcasts, divided into 43,907 shots.	100 hours Dutch television broadcasts and historic footage, divided into 35,766 shots
Queries	67 textual queries (English)	3 sets of queries (Dutch/English): - 36 textual *Archive* queries - 29 multimedia *Future* queries - 72 multimedia *Lab* queries
Automatically generated content metadata	Transcripts (English)	3 kinds of metadata: - Transcripts (Dutch and English) - Low-level feature annotations - Detectors for 57 visual concepts
Manual catalog annotations	*None*	219 broadcast level catalog entries, created by audiovisual archivists.
Visual concept labels	450 visual concepts contained in a unified multimedia thesaurus, manually labelled at the shot level.	*None*

manually judged with respect to 67 queries and 450 visual concepts, making it a particularly well-suited video collection for studying retrieval with visual concepts. The video collection was first released as part of the TRECVID 2005 benchmark evaluation initiative. The initiative released two data sets: a development data set (for learning concept detectors) and a test data set (for evaluating retrieval) [116]. The Broadcast News collection that we define here is based on the development data set.

7.1.1 Collection Overview

Video data The Broadcast News collection consist of 70 hours of English, Chinese, and Arabic news video from October and November of 2004. It has been automatically divided into 43,907 shots by Christian Petersohn at the Fraunhofer (Heinrich Hertz) Institute in Berlin [121]. A shot here is defined as "a series of interrelated consecutive frames taken contiguously by a single camera and representing a continuous action in time and space" [49]. We follow TRECVID practice and use these shots as the predefined division of the video material.

Automatically generated content metadata Each shot in the video collection is associated with automatically created transcripts of the speech track. The transcripts were created by TRECVID using commercial speech recognition software. In

the case of Arabic and Chinese language video, machine translation was used to convert the transcripts to English [116]. To keep the language consistent over the entire video collection, we only use the final English language transcripts.

Queries In total there are 67 queries associated with the Broadcast News collection. Twenty-four of these queries were created by TRECVID query creators, who viewed video data and formulated information needs based on what they could see in the video [116].[2]. These queries consist of English, natural-language, text descriptions and multimedia examples. The remaining 43 queries were created by the LSCOM consortium [110], who developed a series of use cases in cooperation with intelligence analysts, and used these to create queries.[3] These queries consist of English natural language text descriptions, and are not accompanied by multimedia examples. Together, the LSCOM and the TRECVID queries constitute the final set of 67 queries.

Labelled visual concepts The Broadcast News collection has been comprehensively labelled with respect to hundreds of visual concepts. These labels arise from the efforts of the LSCOM and MediaMill [156] consortia, for the purpose of allowing people to develop detectors for performing detector-based search, as described in Chapter 6. They have released sets of 413 and 101 visual concepts respectively,[4] and have labelled the shots in the Broadcast News collections with these concepts. Each concept is accompanied by a short textual description. This description elaborates on the visual elements that should — or should not — be present. For example, the description for the concept detector *storms* is "outdoor scenes of stormy weather, thunderstorms, lightning." It explicitly indicates that video containing lightning and thunderstorms should be tagged as storms. A full list of the concepts can be found in Appendix B. We unify the concepts from the two consortia in a single multimedia thesaurus of 450 concepts, enriched with semantic information from a large-scale ontology, as described in the next section.

7.1.2 A Unified Multimedia Thesaurus

The shot-level concept label annotations created by LSCOM and MediaMill represent a precious resource, as manual labeling constitutes a major effort. However, the visual concepts constitute only a fraction of the concepts that can be expressed

[2] `http://www-nlpir.nist.gov/projects/tv2005/topics/` Relevance judgments on the TRECVID 2005 development set were donated by CMU [191].

[3] `http://www.lscom.org/useCaseQueries/index.html`

[4] The LSCOM lexicon design includes over 2,000 concepts. However, many of these concepts are not accompanied by truth judgments, and here we only consider those concepts that were annotated at the time of writing.

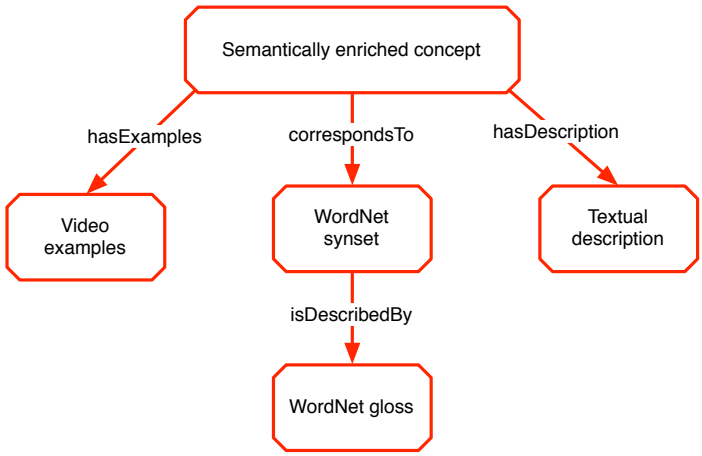


Figure 7.1: Data model for a semantically enriched concept in our multimedia thesaurus.

in the English vocabulary. To compare, the Second Edition of the Oxford English Dictionary contained 597,291 defined terms [139] at the end of December 2009. This richness of vocabulary is a problem for matching searchers' needs to video descriptions [4]. Different terms are used to describe the same video fragment by different users, or by the same user in different contexts. Exploiting ontologies and thesauri to structured terms employed by users can make descriptions more consistent and can aid the searcher in selecting the right term for a visual concept.

To be able to use ontologies in search by visual concept, we link a general-purpose ontology (with over 100,000 concepts) to a set of visual concepts defined specifically for video retrieval (with several hundreds of concepts). This can help, for example, disambiguate between various interpretations or find more general concepts to improve search. As the news domain is broad and can in theory contain any topic, a broad and domain-independent ontology is necessary. As our ontology we use WordNet [107], a lexical database in which nouns, verbs, adjectives, and adverbs are organized into synonym sets (synsets) based on their meanings and use in natural language.

We establish a link between WordNet and the visual concepts from MediaMill and LSCOM. In this way we unify the individual visual concepts in a multimedia thesaurus of semantically enriched concepts. Here each concept is modelled by a set of labelled video examples, a textual description, and a link to a synset in the richly linked WordNet thesaurus, as shown in Figure 7.1. The link is created by humans, who use the textual concept descriptions provided by LSCOM and MediaMill to locate matching synsets in WordNet. The textual descriptions are by no means ex-

haustive, usually consisting of one or two sentences [110, 156], but do contain a significant amount of information about the different kinds of visual content associated with each concept. In the linking process, a human compares the textual descriptions associated with each concept to WordNet glosses, which are short descriptions of the synsets. Each concept is linked to 1–6 synsets, with, at most, two per part of speech (noun, verb, adjective). Concepts describing specific persons that are not present in WordNet are linked as instances of a noun-synset. E.g., *Ariel Sharon* is not present in WordNet and is linked as an instance of the noun-synset "Prime Minister." Each concept was linked to WordNet by two people independently. Initial overlap between the humans was only 65%, and concepts without initial agreements were discussed until agreement was reached.

When concepts in the MediaMill and LSCOM lexicons were linked to the same WordNet synset they were considered to be identical. In this case the annotations from both lexicons were combined. This process resulted in a final multimedia thesaurus of 450 visual concepts, embedded in the rich structure of the WordNet ontology.

7.2 The Archive Footage Collection

Our second video collection, the *Archive Footage* collection, is based on a sample of the audiovisual broadcasts maintained by the Netherlands Institute for Sound and Vision. The broadcasts in this video collection are representative of the archival footage that can be accessed by the searchers that were described in Chapter 3. Like the Broadcast News collection, the video data in the Archive Footage collection was obtained from the TRECVID initiative. It was released as the test data set of the 2008 benchmark release [144].

7.2.1 Collection Overview

Video data While the Broadcast News collection contains only news videos from 2004, the video data in the Archive Footage collection is widely heterogeneous. It contains Dutch documentaries, educational programs, news reports, news magazines, and archival footage, with the oldest program first broadcast in 1927, and the most recent in 2004. In total, the Archive Footage collection encompasses 100 hours of Dutch archived television broadcasts, 219 programs in total. Like the Broadcast News collection, the Archive Footage collection is accompanied by automatically generated shot segmentations provided by the Fraunhofer Institute [121]. In total the programs have been divided into 35,766 shots.

Automatically generated content metadata This consists of transcripts, low-level visual features, and detectors. Each shot in the video collection has been associated with Dutch-language speech recognition transcripts generated by two different automatic systems: one from the University of Twente [68] and the other from the LIMSI laboratory [35]. In addition, Queen Maryi, University of London [16] provided an English-language machine translation of the University of Twente transcripts. The MediaMill group provided low-level feature annotations, and created them using spatio-temporal sampling of interest regions, visual feature extraction, and codebook transform [169]. MediaMill also provided a set of robust detectors for 57 concepts [159]. Each detector is trained using labelled shots from a development collection external to the Archive Footage collection, using the low-level features as input for a Support Vector Machine, in combination with episode-constrained cross-validation [159].

Manual catalog annotations In today's archive, the main source of information used for retrieval is the text from manually created catalog entries that describe each program. We show an example of such an entry in Figure 7.2. As described in Chapter 3, the archive structures its catalog entries using multiple information fields. We associate each broadcast in the Audiovisual Footage collection with its manually created catalog entry, obtained from the Netherlands Institute for Sound and Vision. We divide the fields into three different types, namely: *free text*, natural language descriptions that describe and summarize the content of a program; *tags*, the structured thesaurus terms detailed in Chapter 3 that describe the people, locations, named entities, and subjects in a program; and *technical metadata*, technical information about a program such as identification codes, copyright owners, available formats, and the progra

Three query sets The shots in the Archive Footage collection have been judged with respect to three sets of queries. Two of these sets are based on searches recorded in the transaction logs of the Netherlands Institute for Sound and Vision, while the third query set consists of queries obtained from video retrieval benchmarking efforts, as we will describe in the following sections.

7.2.2 Archive Query Set

Our first set of queries is the *Archive* query set. This query set is designed to reflect information needs of searchers in the audiovisual broadcast archive. The queries and associated relevance judgments in this query set have been taken directly from the transaction logs of the Netherlands Institute for Sound and Vision. While this type of data has been used in other, non-audiovisual, settings to create queries and

Table 7.2: Manual catalog annotations (translated into English) for a broadcast from the Archive Footage collection. The catalog fields are divided into three different types: technical metadata (tech. metadata), free text, and tags.

Field Type	Field Name	Content
Tech. metadata	Title	Noorderlicht — The Image of the Dolphin
Tech. metadata	Broadcast date	1996-11-10
Tech. metadata	Carrier number	HETBEELDVANDE-HRE000038DA.mxf
Tech. metadata	Carrier type	MXF
Tech. metadata	Carrier id	128646
Free text	Summary	Program with reports on scientific topics. In this episode, research by biologist Ken Marten on Oahu, one of the Hawaiian Islands, into the behavior of dolphins.
Free text	Description	Interview with Ken Marten, biologist from environmental organization Earthtrust, about flexible reactions to changing circumstances as a display of intelligence; how dolphins react when they see themselves in a mirror or are confronted with television images; the lack of evidence for the existence of self-awareness in dolphins; the probability that dolphins feel pain when they are killed during the tuna catch. SHOTS: - Oahu: coast; Sealife Park dolphins learn tricks; Marten: watches dolphins through underwater window. ARCHIVE MATERIAL: dolphins under water: playing with rings and tubes of air; view themselves in mirror.
Tags	Genre	Educational; Magazine
Tags	Location	Hawaii
Tags	Person	<no entry>
Tags	Name	<no entry>
Tags	Subject	biology; dolphins; behavioral science; scientific research; intelligence; pain
Tags	Maker	Doornik, Jack van; Feijen, Joyce; Hattum, Rob van; Hermans, Babiche

relevance judgments for textual retrieval experiments [86, 124]; our approach is different because we use *purchase* data rather than click data. In addition, the temporal aspect of the purchase data is unusual, in that it allows us to identify the parts of a video that were purchased as the result of a query. We interpret a purchased video fragment as a fulfilled information need, allowing us to consider the purchase data as relevance judgments in our evaluation [62].

To form an Archive query and associated relevance judgments, we extract information from the transaction logs as follows. We first identify a search session from the archive that has resulted in an order from one of the 219 broadcasts in the Archive Footage collection. We then concatenate the text from the searches in the session to form the text query. We use the ordering data to create relevance judgments at the shot-level. Relevant shots are assigned using the start and end time of the video order; if a shot is contained within an order, then it is considered relevant. The archive queries were created from the set of transaction logs used in Chapter 4, collected between November 18, 2008 and February 1, 2010.

Using the implicit data available in transaction logs to create queries and relevance judgments can be problematic for retrieval experiments. This form of implicit relevance judging does not always clearly delineate the desired visual content. For

Table 7.3: Sample searches and purchases from the transaction logs of searcher actions at the audiovisual archive, used to develop archive queries. Retrieval queries are formed by concatenating consecutive searches in a session; relevant shots are identified using the purchase start and end time within a program.

Queries	Purchase details	Keyframes from order
shots f16 saab airplane shots	*Title*: Zembla: The Defense Orders *Purchase duration*: 13s *Program duration*: 35m 32s	
noorderlicht *on* 1996-11-10	*Title*: Noorderlicht: The Dolphin's Image *Purchase duration*: 25m 13s *Program duration*: 25m 13s	

example, when an entire program is purchased, as in the second example in Table 7.3, all shots within that program are marked as relevant. An alternative to our method of implicit judging would be to use annotators to explicitly annotate individual shot. However, it is difficult to determine the desired visual content — and thus the relevant shots — based on the text of searches alone. Therefore, we create a second set of *Future* queries that is based on more than the text of searches alone by analyzing search sessions in-depth, and accompany them with explicitly created judgments.

7.2.3 Future Query Set

Our *Future* query set has been designed to model searchers' information needs as they might be formulated in the audiovisual archive of tomorrow, an archive equipped with content-based video retrieval capabilities. It is to be expected that the retrieval functionality of the archive will change when the results of multimedia content analysis are included. This will allow users to formulate their queries in new and more diverse ways. We design the Future queries to take advantage of the possibilities offered by state-of-the-art content-based video retrieval systems, such as those evaluated in the TRECVID benchmarks. They incorporate textual statements of visual information need and multimedia examples, and are accompanied by explicitly created shot-level relevance judgments. The Future queries were created on the basis of the transaction log data described in Chapter 3, which was collected from November 18, 2008 to May 15, 2009; due to timing constraints, the Future queries were created using a smaller set of transaction logs than the Archive queries. Candidate queries were created by a query creator from the archive, and two researchers selected the final queries for inclusion in the Future query sets, as follows.

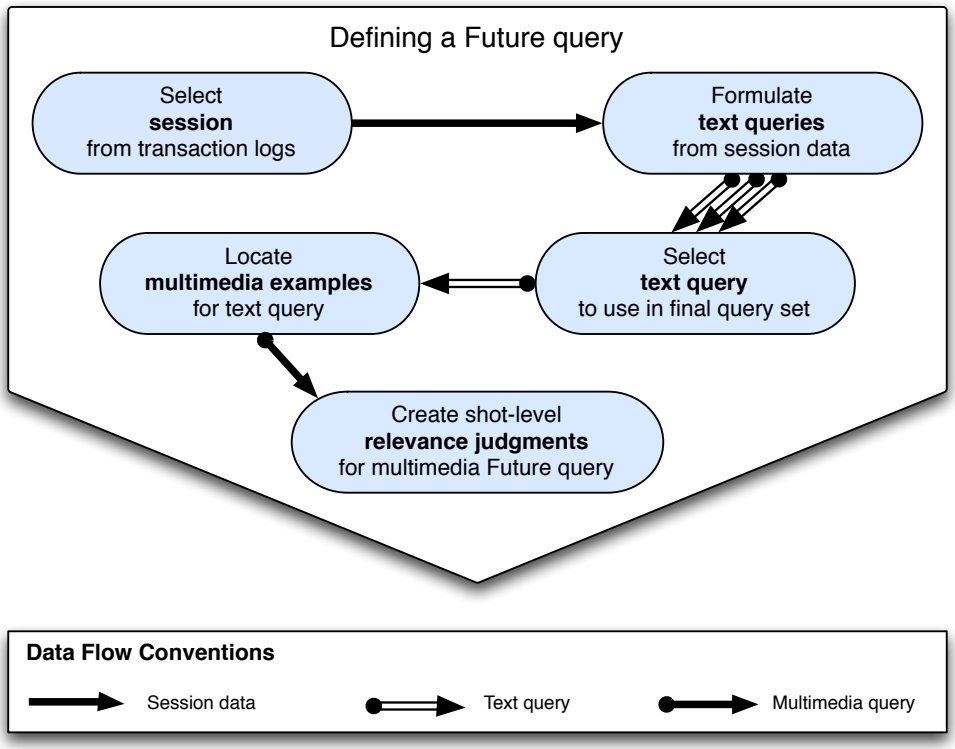


Figure 7.2: Overview of the procedure used to create a query for the *Future* query set, based on transaction logs.

The first step in creating a Future query, as shown in Figure 7.2, was to select a search session in the transaction logs to use as the basis for query development. In this step we selected sessions that resulted in a purchase from a video in the Archive Footage collection, as they were more likely to contain relevant video material than sessions that did not result in a purchase. In total we identified 16 sessions conforming to this criterion. To gain additional sessions, we also chose at random sessions that resulted in the purchase of a video fragment shorter than two minutes in duration from outside of the Archive Footage collection. We chose sessions resulting in purchases of short fragments as it was easier to infer the possible information need of the searcher from these purchases than when longer video fragments were purchased. We selected 8 sessions conforming to this criterion, giving us a total of 24 sessions for query development. For each of the sessions, we extracted the searches, result clicks, and purchasing information from the transaction logs.

In the next step, the session data was used to formulate candidate text queries.

This step was performed by a query creator from the archive, who was asked to analyze the data from a session and develop queries that she felt reflected the underlying visual information need of the broadcast professional. The annotation guidelines were as follows:

1. Scan the session to get an idea of the general information needs of the searcher;
2. View the video fragments that were ordered;
3. Note down the visual information needs that the user may possibly have had; and
4. Rank the noted information needs according to the confidence that they reflect the actual information need of the user.

This resulted in a ranked list of textual queries per session.

Once the query formulation process was completed, text queries were selected for use in the Future query set. This was done by two researchers (including the author) who had knowledge of the Archive Footage collection. We examined the candidate textual queries and selected those that were likely to have relevant examples in the Archive Footage collection. In cases where the queries were considered sufficiently diverse, multiple queries were selected from the same session. In this step the textual queries were also translated from Dutch into English.

To transform a textual Future query into a multimedia Future query, the query selectors associated it with 1–5 video examples from an external collection.

Finally, a group of judges create shot-level relevance judgments on the Archive Footage video data for the Future query. To accomplish this, the judges used an interactive annotation tool [34] to locate relevant and non-relevant shots. Each judge was able to browse through the video using transcript-based search, feature-based search (using online learning), and detector-based search, and associative browsing through the video timeline as described by Snoek et al. [159]. Each judge was given a minimum of half an hour and a maximum of one-and-a-half hours per query to find as many relevant shots as possible. This resulted in the final set of 29 Future queries, for which an overview is given in Figure 7.3.

7.2.4 Lab Query Set

Our *Lab* query set is obtained from existing content-based video retrieval evaluations. Specifically, it incorporates queries from the TRECVID 2007 and 2008 retrieval tasks [146] and the 2008 VideOlympics interactive retrieval showcase [160]. As the video collections used in these initiatives vary from year to year, the queries have relevance judgments on different collections. We performed additional judging to identify relevant shots in the audiovisual broadcasts from the Archive Footage

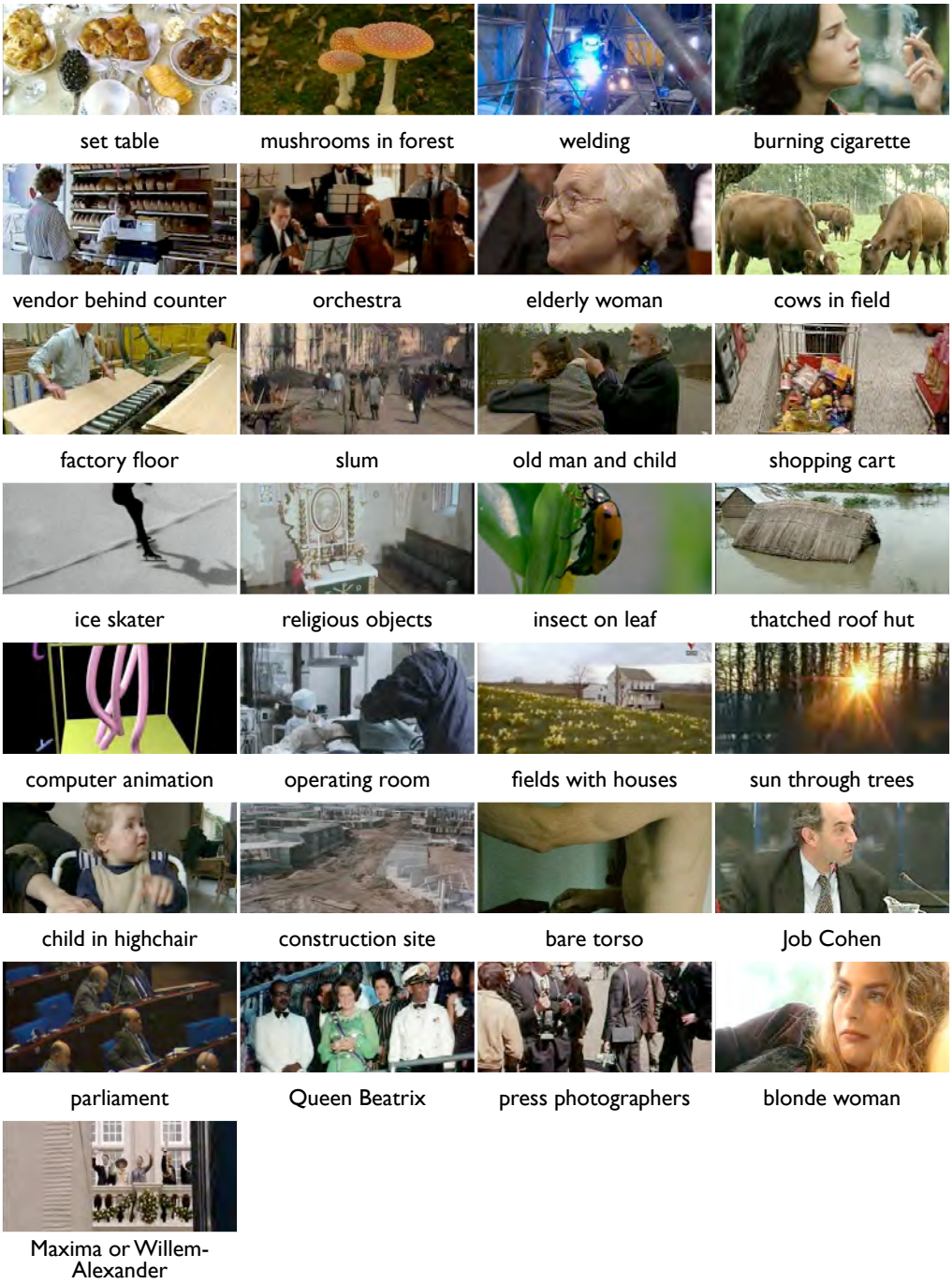


Figure 7.3: Visual overview of the *Future* query set, showing a keyframe from one of the multimedia examples for each of the queries in the set.

collection. This was done using an interactive annotation tool, as described in Section 7.2.3.

7.3 Conclusion

In this chapter we have defined the two video collections that we use as the basis for our experiments throughout the rest of the thesis. The Broadcast News collection is based on Arabic, Chinese, and English news bulletin broadcasts, while the Archive Footage collection is based on a set of Dutch archived broadcasts from the Netherlands Institute for Sound and Vision.

These two collections are suited to our aim of studying how content-based video retrieval methods may be used to help improve search in the audiovisual archive, as they both contain the queries, relevance judgments, and automatically generated content metadata that are essential for performing content-based video retrieval experiments.

The Broadcast News collection is especially well-suited for performing experiments with large numbers of visual concepts, and in Chapter 8 we use it to study methods for assessing automatic concept selection techniques. We also use it in Chapter 9 to investigate the redundancy between mentions of visual items (concepts and queries) in speech transcripts and their appearance in video.

The Archive Footage collection, though not labelled with respect to visual concepts, contains both queries and manually created archive text obtained from a real-world audiovisual archive. This makes it an appropriate testing ground for studying content-based video retrieval in the audiovisual broadcast archive setting. We use the Archive Footage collection as a second video collection in Chapter 9 to investigate the redundancy between mentions of query words in speech transcripts and their appearance in video. The collection also forms the basis for our experimentation in Chapter 10, where we study the extent to which content-based video retrieval can enhance retrieval practice in audiovisual broadcast archives.

Chapter 8

Assessing Concept Selection for Video Retrieval

Having outlined the collections that we will use to perform our retrieval experiments, we will now zoom in on the problem of matching queries to a concept vocabulary. As we described in Section 6.3.1, concept detectors automatically assign estimations of the presence of visual concepts in a shot, for example such concepts as *Sky*, *Airplane*, or *Building*. These concept detectors can then be used for searching; for example, for a user query *Find shots of a field with cows*, the detectors *Animal* and *Grassland* could be used to return shots that are likely to (help) answer the query. Even better would be to use a detector for the concept *Cow*; however, resources for developing concept detectors are limited, and therefore the number of concepts available in a concept vocabulary are limited too. For example, the multimedia thesaurus from the Broadcast News collection described in the previous chapter contains manually assigned labels for 450 concepts, and is thereby representative of the largest concept vocabularies available today [153]. A problem that arises is matching a query to concepts from the concept vocabulary, so that we can identify which detectors to use for search. This brings us to our third main research question,

> **RQ 3** *Given that search by visual concept detector is a valuable method for content-based video retrieval, how can we identify the correct visual concepts to use for a query?*

We term the process of machine-based identification of the visual concepts to use for a query *automatic concept selection*. We saw in Section 6.3.1 that assessing automatic concept selection algorithms is a difficult task. Such algorithms are usually

assessed within the context of end-to-end retrieval system performance [19, 51, 111, 113, 158, 179]. To assess the concept selection task, the scores from the associated concept detectors are used as proxies for the concepts themselves. However, concept detection produces uncertain results, i.e., a shots with a high probability estimate of containing a concept does not necessarily contain that concept. As a result, concept detectors can give unexpected results; for example, in [158] the best concept detector for answering the query *find shots of George Bush entering or leaving a vehicle* was that for the concept *rocket-propelled grenades*, even though the concept *George Bush* was in the concept vocabulary. Furthermore, the overall performance of state-of-the-art concept detectors is constantly improving [151]. Thus a concept detector that works poorly today may perform very well tomorrow. Therefore another approach to assessing the performance of automatic concept selection is necessary.

An alternative to assessing automatic concept selection in the context of a retrieval system is to create an ideal *query-to-concept mapping* for a set of queries [52]. In query-to-concept mapping, a knowledge-intensive method is used to decide which concepts from a given concept vocabulary will be the best to use for retrieval. The concepts selected by a given automatic concept selection algorithm for a query can then be assessed by comparing them to the concepts from the ideal query-to-concept mapping. However, a problem in creating an ideal query-to-concept mapping is identifying which concepts are the correct ones for a query; previous research has shown that humans in isolation typically reach very low agreement on this task, in the range of 15% [21, 52, 113]. Simply asking people individually to identify the correct concepts for a query is therefore not sufficient in creating an ideal query-to-concept mapping.

Therefore we propose the creation of *concept selection benchmarks*, each consisting of a set of ideal query-to-concept mappings, for assessing automatic concept selection. Rather than relying on people individually creating query-to-concept mappings, we investigate two alternative approaches. The first approach is to generate a *collection benchmark* by mining associations between queries and concepts from the annotated shots in a collection that has been manually annotated with both relevance judgments for a set of queries and concept labels from a concepts vocabulary. The second approach is to create a *human benchmark* through explicit human association of concepts to queries, using a focus group setting to arrive a communal decision as to the best concepts for a query. As identifying ideal query-to-concept mappings is a task subject to interpretation, as witnessed by the low human agreement in previous studies, we will compare the query-to-concept mappings made by each of the mapping. Finally, we will use each of the benchmarks to assess the output of automatic concept selection algorithms, in order to develop recommendations as to the application of the concept selection benchmarks for assessing automatic concept selection algorithms. As this is the first time that the two approaches have

been applied to the same set of queries and concepts, we include an analysis of the similarities and differences between the two benchmarks that are produced.

Our study is directed by the following research questions:

CRQ 1 What are the differences in terms of semantics of the query-to-concept mappings produced by collection-based versus human-based development of concept selection benchmarks?

CRQ 2 What is the retrieval performance of the query-to-concept mappings contained in the concept selection benchmarks when applied to retrieval in a collection that has been labelled with visual concepts? In other words, how well do the query-to-concept mappings in the benchmarks work for retrieval when we have perfect detectors?

CRQ 3 How consistently can query-to-concept mappings be reproduced for each of the benchmark creation methods that we describe?

CRQ 4 How successful is each of the two concept selection benchmarks at assessing the comparative performance of automatic concept selection algorithms?

By answering these research questions we hope to attain an increased understanding of both the query-to-concept mapping problem, and of the potential for benchmarks containing ideal query-to-concept mappings to assess automatic concept selection independently of concept detection performance.

8.1 Concept Selection Benchmark Development

We develop two concept selection benchmarks that can be used to evaluate automatic concept selection strategies, based on the two types of non-automatic concept selection approaches identified from previous work in Section 6.3.1. Each concept selection benchmark consists of a set of queries that have been mapped to visual concepts. A given query may be mapped to different concepts by each benchmark, as the methods used to develop the benchmarks are orthogonal, as we will describe.

Collection The development of each concept selection benchmark requires, at minimum, a set of queries and a lexicon of visual concepts that can be used to create query-to-concept mappings. In addition, we require a training set of shots and a test set of shots, where the shots have been labelled with the concepts from the given lexicon of visual concepts, and judged with respect to the given query set used for benchmark development. The training set of shots will be used to generate the collection benchmark, by mining associations between queries and concepts. The test set of shots will be used to perform retrieval experiments with "perfect detectors."

The Broadcast News collection described in the previous chapter is well-suited to our purpose. Recall that the shots in this collection have been manually labelled with respect to our multimedia thesaurus containing 450 visual concepts, and the shots have also been judged with respect to 67 queries. In order to create a training and a testing partition, we split the Broadcast News video data in half chronologically by source, following the TRECVID strategy [116]. We do not split individual news broadcasts, in order to avoid undue influence of the narrative structure of video on our results [172]. The chronological first half of the split is used for training, and the second half is used as our test set. In an additional modification, we removed 1–3 shots per query from the training collection, so that they could be used as query examples in our automatic concept selection algorithm. After these operation, of the 67 queries in the Broadcast News collection, only 50 of them had relevant shots in both the training and the test set. These are the queries that we will use when reporting figures for the concept selection benchmarks.

Retrieval with perfect detectors In this section and in the following sections we will describe retrieval experiments that we performed on the modified test set described above. For each query in the query set, the concepts from the query-to-concept mapping will be used to retrieve shots from the test set. Recall that the test set is manually labelled with concepts (rather than automatically labelled with uncertain concept detectors), therefore we treat these concepts (or more accurately, the concept identifiers), as text. At retrieval time, the concepts for a given query are passed to a text retrieval search engine, and well-established BM25 [129] formula is used to return a ranked list of shots that are likely to be relevant to the query concepts. In the querying process, we do not incorporate the weights available in our concept selection benchmark, as we wish to make as few assumptions as possible about our system. The result list is then evaluated in terms of the average precision evaluation measure that we previously described in Chapter 6, using the query relevance judgements in the test set.

8.1.1 Collection Benchmark

The collection benchmark is created by utilizing explicit judgments about the presence of queries and concepts in shots. The intuition here is that when concepts often occur in relevant shots for a query, but occur less frequently in non-relevant shots, they should be selected as concepts for that query.

The process we use to map a query Q to a set of concepts L_M drawn from a concept lexicon L for the collection benchmark is based on the suggested approach of Hauptmann et al. [57], and can be summarized as follows:

1. Rank all concepts in L by their informativeness as to the relevance or non-relevance of a shot to Q.
2. Discard negatively informative concepts (i.e., concepts whose absence indicates that a shot is relevant to Q), using pointwise mutual information (as we wish to identify the concepts that are most indicative of the presence of a query in a shot, negatively informative concepts are not useful here).
3. Discard concepts with a low informativeness. L_M consists of the remaining set of concepts.

Let us describe each each step in the query-to-concept mapping process in more detail.

In the first step of the process, we employ the information-theoretic notion of mutual information suggested by Hauptmann et al. [57] to determine the informativeness of each concept in L for Q. To do this, we utilize the shot-level query relevance judgments and concept labels available in our training set. Denote relevance of a shot for a given query as R and the presence or absence of a concept in a shot as C. Both R and C are binary random variables. The mutual information between R and C is then defined as:

$$I(R;C) = \sum_{r \in R} \sum_{c \in C} P(r,c) \log \frac{P(r,c)}{P(r)P(c)} \tag{8.1}$$

with the indicator functions $r \in \{\texttt{relevance}, \texttt{non}-\texttt{relevance}\}$, $c \in \{\texttt{presence}, \texttt{absence}\}$; estimates are derived from a shot collection associated with concept judgments [101]. Concepts are ranked according to how much $I(R;C)$ reduces the entropy of R using maximum likelihood estimation (MLE). The estimation is given by $p(\cdot) = \frac{n(\cdot)}{|\mathcal{C}|}$ where $n(\cdot)$ is the number of occurrences in the collection $\mathcal{C}$ and $|\mathcal{C}|$ is the total number of documents (shots) in the collection. The ranked list produced by this method assigns high scores to both positively and negatively informative concepts for a query.

Next, we identify and remove negatively informative concepts from the ranking. We identify negatively correlated concepts using *pointwise mutual information*:

$$I(r;c) = \log \frac{P(r,c)}{P(r)P(c)}. \tag{8.2}$$

If $I(\texttt{absence};\texttt{relevance}) > I(\texttt{presence};\texttt{relevance})$ we discard that concept from the query-to-concept mapping.

What remains is a scored, ranked list of all positively informative concepts for each query. Hauptmann et al. [57] suggest the use of a cut-off threshold of 1% reduction in entropy of R to remove concepts with a low informativeness from the

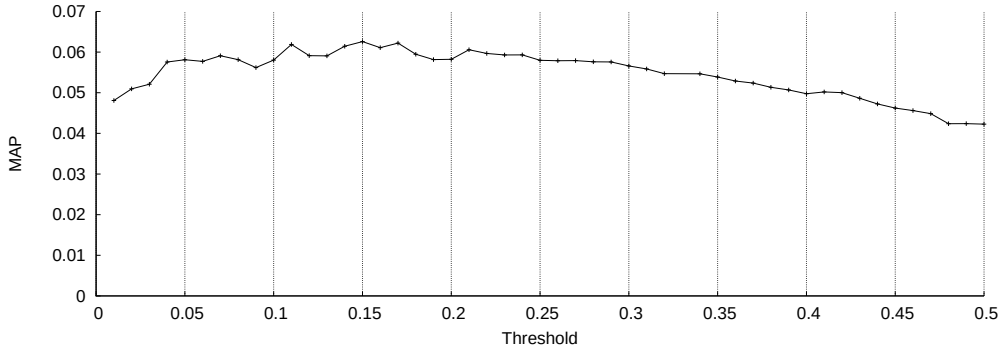


Figure 8.1: The effect on retrieval performance on the test set of changing the threshold value at which concepts are included in the collection benchmark. As the threshold value increases, less concepts are included in the benchmark.

list. When we applied this threshold, we found that the number of concepts in the query-to-concept mappings were still quite large, and it was not clear whether these mappings would be useful for assessing concept selection techniques. To test this, we performed retrieval experiments on the test collection with different threshold settings, shown in Figure 8.1. We found that retrieval performance using the collection benchmark mappings gained a relative improvement of 31%, from a MAP of 0.048 to 0.063, by using a threshold of 0.15 rather than a threshold of 0.01. Therefore, we use the threshold value of 0.15 to remove concepts with low informativeness from our ranked list. The remaining concepts form the concepts L_M in the query-to-concept mapping for Q.

8.1.2 Human Benchmark

The human benchmark is created by asking human judges to select the concepts that are appropriate for each query. The intuition here is that humans know which concepts should be used to answer a given information need.

People can have a wide range of associations with a concept, depending on context and personal characteristics. Nevertheless, there exists a common understanding of concepts that is socially constructed and allows people to communicate [97, 137]. The goal of the human benchmark is to capture this common understanding, as opposed to the wider range of individual associations [57, 113]. Therefore, rather than using individual human assessors to map queries to concepts, we create the human benchmark using focus group experiments, where multiple individuals must reach consensus as to which concepts are appropriate for a query.

We conducted two focus group experiments, each following the same procedure.

Table 8.1: The effect on (1) retrieval performance, and (2) average number of concepts mapped per query, when changing the number of votes required for a query-to-concept mapping to be included in the human benchmark. As the focus group consisted of four participants for group 1 and three participants for group 2, the maximum number of possible votes a query-to-concept mapping could obtain was four.

	Vote threshold		
Measure	*At least 1 vote*	*More than 1 vote*	*Unanimous*
Retrieval MAP	0.059	0.057	0.059
Concepts per query	8.4	6.6	3.5

The first was conducted in January 2008 and the second in March 2008, with respectively 3 and 4 undergraduate students. Both studies consisted of several sessions held over a two week period. None of the subjects had prior experience retrieving video with detectors. To facilitate comparison between the two focus group studies, we included 30 overlapping queries.

Each experiment consisted of a *familiarization* phase and a *mapping* phase. The familiarization phase was designed to make the subjects accustomed to the concept vocabulary available in the multimedia thesaurus. A total of 450 flash cards, each printed with the name and textual description of a concept from L, were given to the group. The subjects were asked to collaboratively organize the cards so that relevant concepts for a query could be identified quickly.

In the mapping phase, focus group participants collaborated to create a final query-to-concept mapping. This phase consisted of the following steps for each query: In the first step, participants were given 8 queries at a time and asked to note down, in isolation, the concepts that would help them find relevant shots in a video collection.

In the second step, the focus group moderator called on each participant in turn, and noted down their candidate query-to-concept mapping for a given query on a flipboard. All concepts were written down, even if the participant had changed their mind about the usefulness of a concept.

Finally, all participants voted on which concepts they thought would help them find relevant shots, and would therefore be the most appropriate for including the final query-to-concept mapping. During this group process discussion was encouraged. As a result, concepts in the list might receive no votes. These concepts were removed from the list. Participants were also asked to unanimously decide which single concept would be best for retrieving relevant shots. The final query-to-concept mapping consisted of all concepts that gained at least one vote.

Only 47% of the concepts in the final mapping were unanimously voted as appopriate for inclusion in the benchmark. As we found that increasing the threshold

value for keeping concepts in the query-to-concept mappings improved the retrieval performance of the collection benchmark, we investigated whether a similar phenomenon could be observed for the human benchmark. Therefore we performed three retrieval experiments where we used the number of votes as a threshold. In the first experiment we included all concepts with at least one vote, in the second all concepts with more than one vote, and in the third only those concepts with a unanimous decision (three or four votes, depending on the number of people in the focus group). As can be seen in Table 8.1, there was no appreciable difference in retrieval performance between the three approaches. To enable us to test our ranking-based measure in assessing automatic concept selection approaches later on in the chapter, we decided to include all concepts with at least one vote in the final set of concepts L_M. In this way, the number of votes for a concepts can be used to rank them within the mapping.

8.2 Analysis of the Concept Selection Benchmarks

Over the 50 queries mapped by both concept selection benchmarks, the collection benchmark mapped 6.7 concepts per query and the human benchmark mapped 5.6 concepts per query. Out of the available 450 concepts, the collection benchmark contains 125 unique mapped concepts, while the human benchmark contains 163 unique mapped concepts. Together, the benchmarks used 203 of the available 450 concepts for the 50 overlapping queries.

Set agreement In our analysis of the concept selection benchmarks, and later in our assessment of automatic concept selection algorithms, we will compare the overlap between different sets of concepts for a query. To compare this overlap, we use *set agreement* (SA), which is also known as the positive proportion of specific agreement or the balanced F-measure [66]. Consider two sets of concepts L_1 and L_2 for a given query Q. The set agreement is then given by

$$\mathrm{SA}(L_1, L_2) = \frac{2 \times |L_1 \cup L_2|}{2 \times |L_1 \cup L_2| + |L_1 \cap L_2|}. \tag{8.3}$$

Set agreement is equal to 1 when $L_1 = L_2$, and 0 when $L_1 \cap L_2 = \emptyset$.

8.2.1 Semantic Overlap Between the Benchmarks

To compare the semantic overlap captured by the two concept selection benchmarks we calculated the per-query set agreement between the two concept selection benchmarks for the 50 queries that were contained in the benchmark, shown in Figure 8.2.

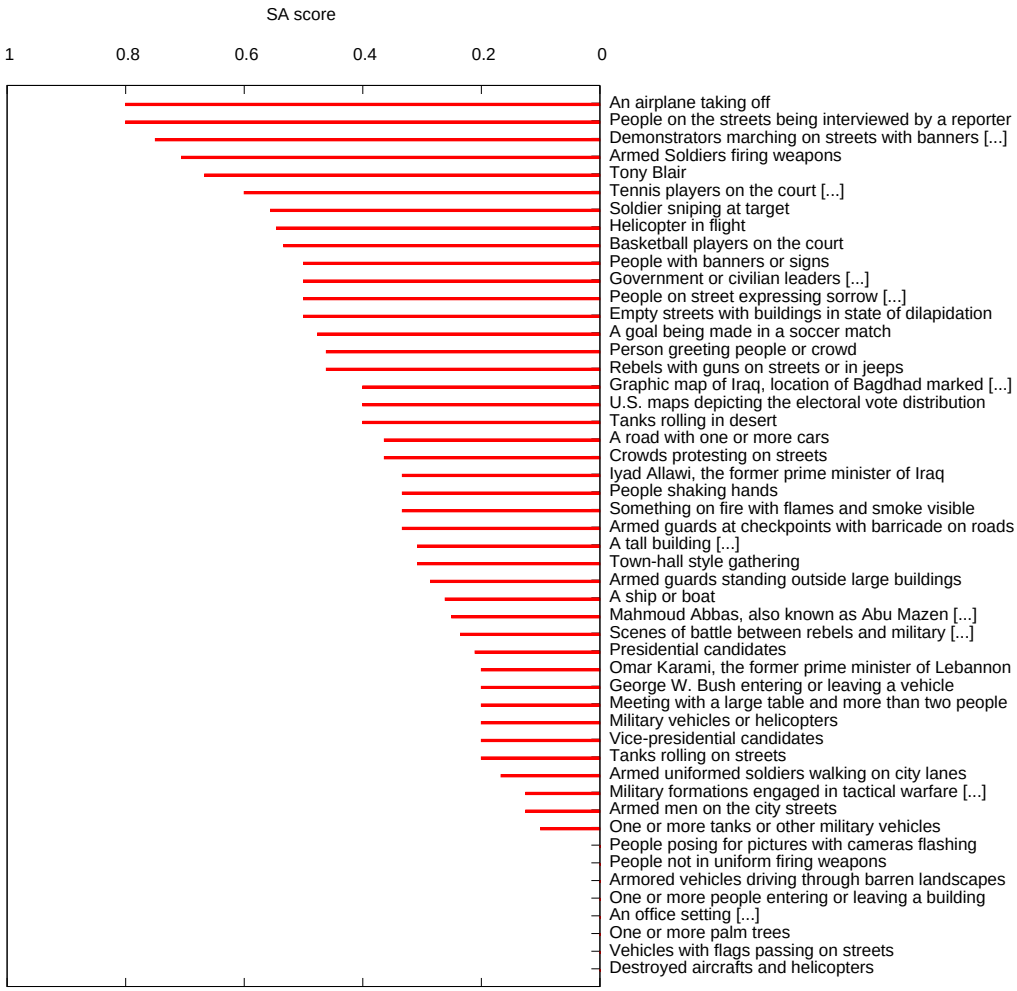


Figure 8.2: Per-query set agreement between the query-to-concept mappings created by the collection benchmark and the human benchmark. Queries are sorted by set agreement score.

The average SA was 0.34. For no query was there perfect overlap between the query-to-concept mappings of the two benchmarks (where SA would be equal to 1). For seven queries there was no overlap in the mappings (where SA was equal to 0). To gain more insight into the similarities and differences between the benchmarks, we analyze some of the queries with the highest and the lowest SA scores. The mappings for these examples are given in Table 8.2.

Queries with high overlap Beginning with the queries with high agreement between the concept selection benchmarks, for the query *An airplane taking off* the collection query-to-concept mapping contained all of the concepts that the human benchmark mapping did. In addition, the collection benchmark included three extra concepts, *Aircraft*, *Vehicle*, and *Airplane Landing*. The first two of these concepts are generalizations of the *Airplane* concept that is mapped by both benchmarks. *Airplane Landing* is interesting in that is the opposite of the action of taking off that is specified in the query. To gain an understanding of this, we looked at the relevant shots in the training collection that were used to generate the collection benchmark query-to-concept mapping. In total there were seven shots marked relevant to the query, three of these were labelled with the concept *Airplane Landing*. We attribute this to the visual similarity of an airplane take-off and an airplane landing - concept labelling is usually done on the basis of viewing individual keyframes rather than motion sequences, and thus it is difficult for labellers to distinguish between the take-off and landing actions of an airplane.

Turning to the second query, *People on the streets being interviewed by a reporter*, here the human benchmark query-to-concept mapping contained all of the concepts that the collection benchmark did, and in addition also the concept *Reporters*. This could be due to a misinterpretation effect; note that the reporter is not the subject of the query, and need not necessarily be visible for a shot to be relevant. Once again, we examined the relevant shots used to generate the collection benchmark mapping. In this case, there was a total of 65 relevant shots in the collection, but only 11 of them were labelled as containing *Reporters*. In contrast, 60 of the relevant shots were labelled with the concept *Interview on Location* and 54 were labelled with the concept *Microphone*.

Queries with no overlap Now we turn to the queries with no overlap between the query-to-concept mappings of the two concept selection benchmarks. One of these queries is for *One or more palm trees*. Here the mapping from the human benchmark contains concepts one might typically think of when imagining videos containing palm trees; sunny *Tropical Settings*, with perhaps a *Beach* in the foreground or an oasis in the *Desert*. The mapping from the concept selection benchmark, however, is tuned to the news domain, in which such idyllic settings are not featured dominantly and armed conflicts are more common. In an examination of the training collection, we found that the relevant shots used to generate the mapping for the collection benchmark were almost all about a battle taking place in the Iraqi city of Fallujah, a location that contained many palm trees in the background of the shots. As a consequence, concepts such as *Weapons* and *Shooting* attained high mutual information scores and were included in the mapping.

Turning to the query *Vehicles with flags passing on streets*, the query-to-concept

Table 8.2: Example query-to-concept mappings in the collection benchmark and the human benchmark, selected according to the set agreement (SA) between the concepts mapped by each benchmark for a query.

Query text	*Collection benchmark*	*Human benchmark*	*SA*
An airplane taking off	Airport, Airport Or Airfield, Aircraft, Airplane Flying, Airplane Takeoff, Airplane Landing, Airplane, Vehicle, Runway	Airplane Takeoff, Runway, Airport, Airplane, Airport Or Airfield, Airplane Flying	0.80
People on the streets being interviewed by a reporter	Interview On Location, Microphones	Interview On Location, Reporters, Microphones	0.80
Demonstrators marching on streets with banners [. . .]	Protesters, People Marching, Demonstration Or Protest, Crowd, Urban	Demonstration Or Protest, People Marching, Protesters	0.75

(a) Queries with highest set agreement

Query text	*Collection benchmark*	*Human benchmark*	*SA*
One or more palm trees	Weapons, Fire Weapon, Shooting, Outdoor, Daytime Outdoor, Sky	Tropical Settings, Tree, Beach, Desert, Vegetation, Forest	0.00
Vehicles with flags passing on streets	Pickup Truck, Police Security Personnel, Truck, Ground Vehicles, Car, Vehicle, Road, Sky, People Walking, Adobehouses, Dirt Gravel Road, Scene Text, Demonstration Or Protest	Military Ground Vehicle, Flag Usa, Non Us National Flags, Flag	0.00
Destroyed aircrafts and helicopters	Ruins, Demonstration Or Protest, Entertainment	Airplane Crash, Emergency Vehicles, Helicopters, Aircraft, Fighter Combat	0.00

(b) Queries with no set agreement

mapping from the human benchmark contains three types of *Flag* concepts, while mapping from the collection benchmark contained not a single *Flag* concept. Once again, we turned to the training data to examine why this was the case, and found that only two relevant shots were used to generate the mapping for this query. In both of these shots, the flags on the vehicles were small and easy to miss, and neither of the shots had been labelled with any *Flag* concepts.

Finally we examine the query *Destroyed aircrafts and helicopters*, where the mapping from the concept benchmark, surprisingly, does not contain the concepts *Airplane Crash*, *Helicopter*, or *Aircraft*. An examination of the shots that were used to generate this mapping yielded a similar finding as to the last query: there were only two relevant shots in the training data, and neither contained any concept labels that might indicate that a destroyed aircraft was being shown in the video signal. Upon viewing the shots (which were two adjacent shots in the same news broadcast), we found that they depicted archival footage of the crashed aircraft of Senator Ted Kennedy in 1964. The footage was in black and white and of low quality, making it difficult to understand what was being shown in the video signal without listening

to the audio, and therefore resulting in incorrect labeling for visual concepts.

Summary Our analysis of the differences between the concepts mapped by the two concept selection benchmarks to individual example queries has yielded some insights into the causes of differences between the mappings contained in the two benchmarks.

Specificity requirements The concepts mapped to an individual query can vary in their specificity — one benchmark may map the same concepts to a query as the other, but also introduce additional, less specific concepts.

Collection noise Inaccuracies in concept labels assigned to the shots used to generate the query-to-concept mappings of the collection benchmark can introduce irrelevant concepts into the benchmark, especially when there are very few shots used to generate the mapping.

Domain mismatch The relevant shots that are available in a certain domain may not be the same kinds of shots that a human expects to receive as a result for a given query.

Query interpretation The focus group participants creating the human benchmark may interpret the predicate of a query as a required part of the query.

8.2.2 Retrieval Performance

Recall that when we were developing our concept selection benchmarks, we determined the cut-off point for deciding which concepts to include in query-to-concept mappings by applying the mappings to retrieval on our set of test shots. Using the final threshold value of 0.15 resulted in a MAP score of 0.063 for the collection benchmark, and for the human benchmark including all concepts with at least one vote resulted in a MAP score of 0.059. In this section we compare the retrieval performance of the two benchmarks at the level of individual queries.

Figure 8.3 shows the per-query retrieval performance improvement of the collection benchmark as compared to the human benchmark, in terms of average precision. In addition, the absolute retrieval performance scores for each query are shown. For comparison purposes, we say that one score is better than another when the change in average precision is more than 0.01, as we are unable to perform significance tests for individual queries. Using this threshold, the collection benchmark query-to-concept mappings gave better results than those of the human benchmark for 15 queries, and the human benchmark mappings gave the best performance for 8 queries. The remaining 27 queries performed equally well using both benchmarks. However, for those 27 queries, only 15 gave an average precision of more than 0.01.

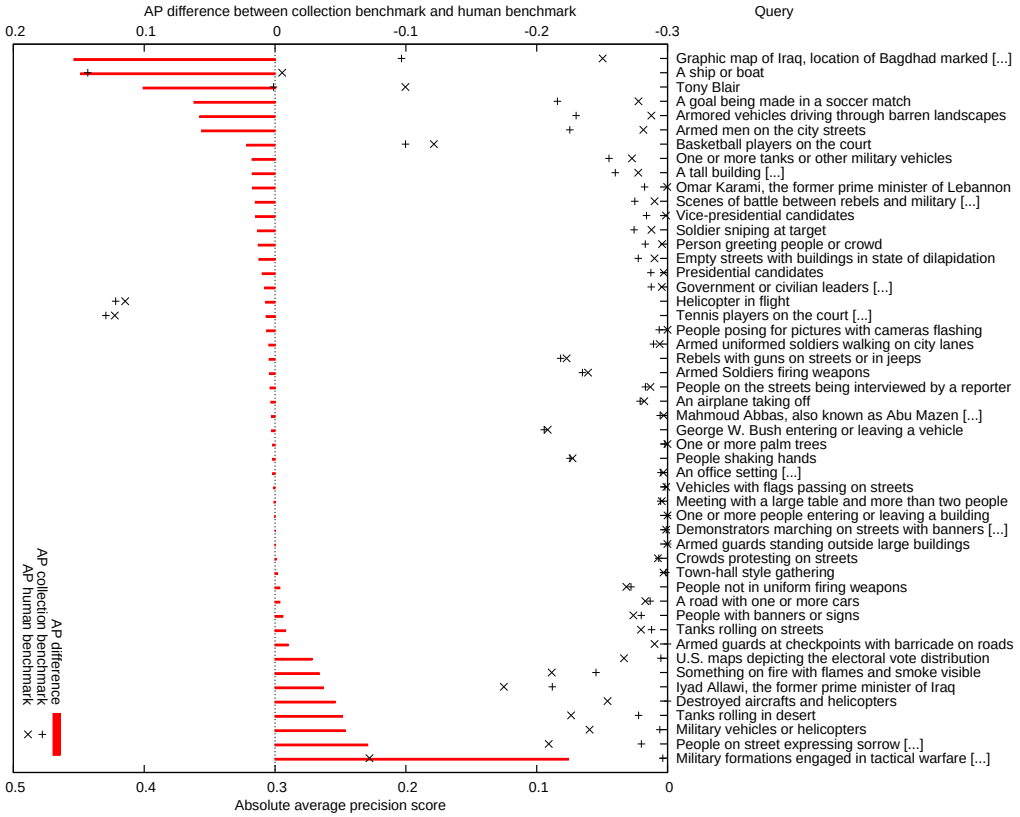


Figure 8.3: Per-query difference in AP scores between human and collection benchmark, ordered by difference in AP score. The queries that perform best using the query-to-concept mappings from the collection benchmark are shown at the top of the graph, while queries that perform best using the human benchmark are shown at the bottom of the graph.

The remaining 12 queries we term "difficult" for detector-based search, as even when we are using "perfect" detectors with benchmarked query-to-concept mappings they result in very low performance.

Let us once again look at some example query-to-concept mappings, shown in Table 8.3, this time selected according to retrieval performance.

Queries where the collection benchmark did better For our first example query, *Graphic map of Iraq, location of Bagdhad* [sic] *marked*, the collection benchmark mapping consists solely of the two concepts *Graphical Map* and *Graphics*. The

Table 8.3: Examples of query-to-concept mappings in the collection benchmark and the human benchmark, selected according to the change in AP when applied to detector-based search in a retrieval system with "perfect" detectors.

	Collection benchmark		**Human benchmark**	
Query text	*Concept mapping*	*AP*	*Concept mapping*	*AP*
Graphic map of Iraq, location of Bagdhad marked [. . .]	Graphical Map, Graphics	0.203	Graphical Map, Cityscape, Urban	0.050
A ship or boat	Boat, Waterscape, Harbors, Waterways, Ship, Vehicle, Lakes, Outdoor, Daytime Outdoor	0.443	Harbors, Freighter, Boat, Shipyards, Sailboat, Barge, Cigar Boats, Ship, Rowboat, Tugboat, Houseboat, River, Canal, Oceans	0.295

(a) Queries where the collection benchmark query-to-concept mappings gave the greatest increase in retrieval performance

	Collection benchmark		**Human benchmark**	
Query text	*Concept mapping*	*AP*	*Concept mapping*	*AP*
Military formations engaged in tactical warfare [. . .]	Adobehouses, Ground Combat, Rifles, Armed Person, People Marching, Soldiers, Military Personnel, Weapons, Violence, Building, People Walking, Dirt Gravel Road, Group	0.004	Military Personnel, Parade, Military Ground Vehicle	0.228
People on street expressing sorrow [. . .]	Protesters, Apartment Complex, People Marching, Residential Buildings, Building, Crowd, Urban, People Walking, Windows, Demonstration Or Protest	0.020	People Crying, Funeral, Demonstration Or Protest, People Marching, Protesters, Crowd	0.091

(b) Queries where the human benchmark query-to-concept mappings gave the greatest increase in retrieval performance

	Collection benchmark		**Human benchmark**	
Query text	*Concept mapping*	*AP*	*Concept mapping*	*AP*
Tennis players on the court [. . .]	Tennis Game, Sport Games, Athlete, Grandstands Bleachers, People Walking, Running	0.429	Tennis Game, Sport Games, Running, Daytime Outdoor	0.423
Mahmoud Abbas, also known as Abu Mazen [. . .]	Government Leader	0.006	Head Of State, Government Leader, Male News Subject, Muslims, Yasser Arafat, Non Us National Flags, Ariel Sharon	0.003

(c) Queries where the mappings of both benchmarks gave approximately equal retrieval performance

human benchmark mapping includes the concepts *Cityscape* and *Urban*. The inclusion of these two concepts introduces noise into the result list in the form of shots that are labelled with them, but not with the *Graphical Map* concept, reducing the retrieval performance. Going back to the original creation of the human benchmark,

we found that the focus group members who created the query-to-concept mappings disagreed about the inclusion of the two noisy concepts, which received only one vote each, while they unanimously agreed that the *Graphical Map* concept should be included.

For our other example query, *A ship or boat*, the mappings of both benchmarks include primarily nautical concepts, and it seems surprising that the collection benchmark mappings do so much better. We manually inspected the first ten results for each benchmark mapping, and found that all of them were relevant, i.e., they contained ships or boats. However, for the results from the human benchmark mapping, the first, eighth, and ninth results were not judged relevant in the ground truth, while for the collection benchmark only the seventh result was not judged relevant. AP strongly penalized highly ranked not-relevant results, therefore these missing relevance judgments resulted in a lower score for the human benchmark mapping.

Queries where the human benchmark did better The query where the human benchmark mapping improves the most over the collection benchmark mapping is *Military formations engaged in tactical warfare [...]*. The collection benchmark mapping includes concepts without an obvious semantic relationship to the query, such as *Adobehouses* and *Dirt Gravel Road*. These concepts introduced not-relevant results into the result list. We investigated why the mapping was so unfocused for this query by viewing the shots that were used to generate the collection benchmark mapping for this query; only two shots, adjacent in the same news broadcast, had been judged relevant in the training set. Coincidentally these shots contained adobe buildings and a gravel road, and consequently these concepts were included in the mapping. This reinforces our observation in the previous section that noise is introduced into the collection benchmark when there are few shots used to generate the mapping.

Turning to the query *People on street expressing sorrow*, our analysis here showed the same phenomenon as for the previous query; concepts without an obvious strong semantic relationship (such as *Windows* and *Building*) were included in the mapping, introducing noise into the results. Once again, an investigation of the shots used to generate the collection benchmark mapping showed that only two were used, introducing concepts that occurred coincidentally in these shots into the mapping.

Queries where mappings from both benchmarks do equally well (or poorly) The final two example queries in Table 8.3 were chosen such that the first is the query where both benchmark mappings attained equally the highest performance, and the second is one of the 12 queries where mappings from both benchmarks gain a score of close to 0. For the query *Tennis players on the court [...]*, the mappings from both benchmarks include only semantically related concepts in the query, resulting

in high performance. For the query *Mahmoud Abbas, also known as Abu Mazen [...]* , the mappings from both benchmarks include similarly semantically related concepts. However, there is no concept specific to the person *Mahmoud Abbas*. The nearest concept, *Government Leader*, occurs in 1,983 shots, in other words, in 8.6% of the test set. However, there are only 33 relevant shots for the query in the collection, and thus the discriminative power of the concepts in our multimedia thesaurus is not sufficient for this query. Presumably this problem could be solved by increasing the size of the multimedia thesaurus; [57] suggest that a concept lexicon of 5,000 results should be sufficient to provide acceptable retrieval accuracy for a large range of queries, even if these concepts are labelled using imperfect detectors.

Summary Our analysis of the retrieval performance of the mappings of our two benchmarks on individual queries showed that, when the mappings from the collection benchmark did much better than the mappings from the human benchmark, this was (in one example at least) because the human benchmark introduced noisy concepts, the inclusion of which focus group participants did not unanimously agree on. This could have been resolved by using a higher agreement threshold for including concepts in the query-to-concept mappings. When the human benchmark did much better than the collection benchmark mappings, for both of the examples that we studied this was due to only a few shots being available for generating the collection benchmark mappings, introducing noise. When neither benchmark did well, this was because the relevant shots were difficult to find using the available concepts in the multimedia thesaurus. We expect that this could be remedied by increasing the number of concepts available in the multimedia thesaurus.

8.2.3 Reproducibility

In this section we investigate the reproducibility of the two concept selection benchmarks, i.e., how consistently can the query-to-concept mappings of each concept selection benchmark be recreated between experiments?

Human Agreement Reproducibility of the human benchmark is assessed by analyzing agreement between user studies 1 and 2 for the 30 queries annotated by both groups. When choosing the best concept, the groups agreed on 80% of the queries. In cases where the two groups did not agree on the best concept for a query, the best concept of one group was still identified as a suitable concept for the query by the other group (just not the most suitable one).

On average, group 1 considered 8 concepts per query relevant, while group 2 selected 5 concepts per query. Group 2 both identified fewer concepts during the individual selection round, and removed more concepts during the voting round. This

difference is attributed to: (1) group size—the group in study 1 had four members, while the group in study 2 only had three, and (2) the physical layout of the concept hierarchy—in study 1, subjects ensured all concepts were visible during the selection process, while in study 2 subjects created a layout where similar concepts overlapped.

Because of the difference in the number of concepts selected across groups, we compare the two sets of selected concepts using asymmetric set difference. The group in study 1 identified 78% of the concepts that were selected by the group in study 2. Group 2 found 34% of the concepts found by group 1. This reflects our previous observation that group 1 selected more concepts per query than group 2 did.

Varying Labelled Training Set Size The collection benchmark may be reproduced by anyone with the annotated video material. But is it reproducible when we vary the number of shots available for generating the benchmark? If less concept labels and less truth judgments for queries are available, will we then get the same sets of mapped concepts? To assess this, we generated the concept selection benchmark using the first 10% of the training set, once again selected chronologically and constrained by episode. We compare the resulting concept mappings to the concept mappings created when using the first 20%, 30%, ..., 100% of the training set and evaluate this against both the (final) collection benchmark, and the human benchmark using set agreement. We restrict our analysis to the 22 queries with relevant shots in the first 10% of the training set.

Figure 8.4 shows the set agreement as the size of the training set is increased. The agreement with the human benchmark slowly increases from 0.30 to 0.39, depending on the amount of training data. The agreement with the collection benchmark converges to 1 (when the training set is the same as the training set used for the final collection benchmark), but starts off with an agreement of 0.5. This variation in set agreement as the collection size changes indicates that the benchmark is sensitive to the size of the collection used to generate the data.

Summary Our analysis of the differences between the query-to-concept mappings created by the human benchmark indicated that it is difficult to reproduce exact query mappings between focus group experiments, as one focus group mapped more concepts per query than the other. However, agreement on the best concept per query was high at 80%, and set overlap for those concepts mapped by focus group 2 (which mapped less concepts on average) was also high at 78%. The benchmark mapping, on the other hand, can be reproduced perfectly by anyone with access to the same labelled collection. However, this query-to-concept mappings of this benchmark become less reproducible as the size of the training set used to generate them

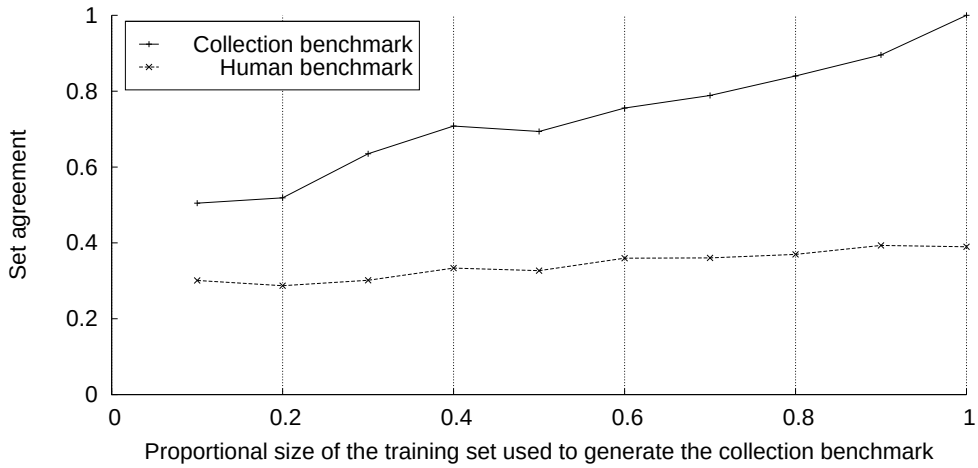


Figure 8.4: Reproducibility of the collection benchmark in generating consistent concept mappings across different proportions of the training set of 20,745 shots. Reproducibility is shown in terms of set agreement, and compared to both the final collection benchmark, and the human benchmark.

decreases.

8.3 Using the Benchmarks to Assess Concept Selection

In this section we examine assessment measures for predicting whether a set of automatically selected concepts will perform well for retrieval. We first summarize our measures for assessing automatic concept selection algorithms using the benchmarks. We then examine the predictive power of the assessment measures for assessing automatic concept selection algorithms for detector-based search, with perfect detectors.

8.3.1 Assessment Measures

Each concept selection benchmark contains query-to-concept mappings that, according to a given benchmark development method, have been defined as the ideal query-to-concept mapping. Therefore the closer the query-to-concept mappings of an automatic concept selection algorithm to the mappings in a benchmark, the closer that the algorithm comes to producing an ideal mapping. Here we define three assessment measures to determine the similarity of a candidate set of selected concepts,

L_S, to the ideal benchmark concept mapping, L_M. The assessment measures take increasing amounts of ranking information from L_M and L_S into account.

SA Set agreement, the measure that we have been using throughout this chapter to compare overlap between query-to-concept mappings. This is a simple set-based measure, that takes into account only the binary absence and presence of concepts in L_M and L_S into account. The maximum possible value is 1 (perfect overlap), the minimum possible value is 0 (no overlap).

MAP Mean average precision, the measure suggested by Hauff et al. [52] for assessing automatic concept selection. The score is based on an analogy to document retrieval: the concepts in L_M are considered to be relevant documents, and the concepts in L_S are considered to be a ranked list of results. MAP is then calculated as described in Chapter 6. This measure takes the ranking of L_S into account, but not the ranking of L_M. The maximum possible value is 1 (all concepts in L_M ranked at the top of L_S, the minimum possible value is 0 (none of the concepts in L_M are contained in L_S).

RC Rank correlation, the measure used to measure the similarity of two rankings in Chapter 4. Each set of concepts is considered as a ranked list, then the similarity between the two rankings is calculated with Kendall's τ, described in Section 4.2.1. Rank correlation requires two sets of rankings for the same items, therefore we add any concept present only in one list to the bottom of the other list. The maximum possible value is 1 (L_M and L_S both contain exactly the same ranked list of concepts), the minimum possible value is -1 (L_M and L_S are perfectly inverse in their ranking order).

8.3.2 Test Case: Assessing Concept Selection

To assess the assessment measures and benchmark for the task of concept selection, we implement three automatic concept selection algorithms from the three families of algorithms identified from the literature. Our aim here is to review the ability of a concept selection benchmark to predict concept selection performance, rather than extensively investigate the merits of individual concept selection approaches. Therefore we do not test variations of concept selection algorithms within the three families.

We perform automatic concept selection for all 50 test queries. We assess the scores of the concept selection algorithms against both concept selection benchmarks, and analyze the predictive power of the concept selection benchmarks for concept-based video retrieval. We do this by contrasting the concept selection benchmark scores with the retrieval performance of the concept selection algorithms, using the

Table 8.4: Mean assessment scores of the three automatic concept selection algorithms — text matching (TM), ontology querying (OQ), and visual querying (VQ) — using the two concept selection benchmarks, over 50 queries. The automatic concept selection algorithm with the highest mean score per combination of benchmark and assessment measure is highlighted in bold

	Collection benchmark			**Human benchmark**		
Assessment measure	*TM*	*OQ*	*VQ*	*TM*	*OQ*	*VQ*
SA	0.178	0.144	**0.443**	**0.224**	0.151	0.138
MAP	0.212	0.153	**0.499**	**0.435**	0.321	0.119
RC	-0.180	-0.331	**0.111**	**0.026**	-0.140	-0.338

retrieval setup of the 'oracle' retrieval system with perfect detectors outlined in Section 8.2.2.

Automatic Concept Selection Algorithms We implement three concept selection algorithms, based on the three families of *text matching*, *ontology querying*, and *visual querying* [153, 158]. For the text matching algorithm, we follow common practice and match query text to a concept's textual description (after normalization and stop word removal) [2, 158, 179]. Concepts are ranked according to their similarity to the original query using the vector space model [132]. All returned concepts are added to the concept selection set. The ontology querying algorithm assigns query terms to nouns in WordNet [107], following [158]. The query nouns are related to concepts, which are also assigned to WordNet nouns. Concepts are assigned scores using Resnik's measure of information content [126]. To prevent this algorithm from returning all concepts in the lexicon we only select concepts with an information content higher than 5. Finally, for the visual querying algorithm, we used the multimedia examples from each query to identify concepts to use for search. As there are no concept detection scores in the collection benchmark, for this method we simply select all concepts that are labelled in the query examples.

Benchmark Scores An overview of the spread of assessment scores produced using the three assessment measures is shown in Figure 8.5. The mean assessment scores are given in Table 8.4. Interestingly, the collection benchmark consistently assigns the highest scores to the visual querying algorithm, while the collection benchmark assigns the highest scores to the text matching algorithm. In other words, the visual querying algorithm produces concepts that are closest to those mapped by the collection benchmark, while the text matching algorithm produces concepts that are the closest to those mapped by the human benchmark. However, no clear picture occurs as to which concept selection algorithm will produce the best retrieval scores.

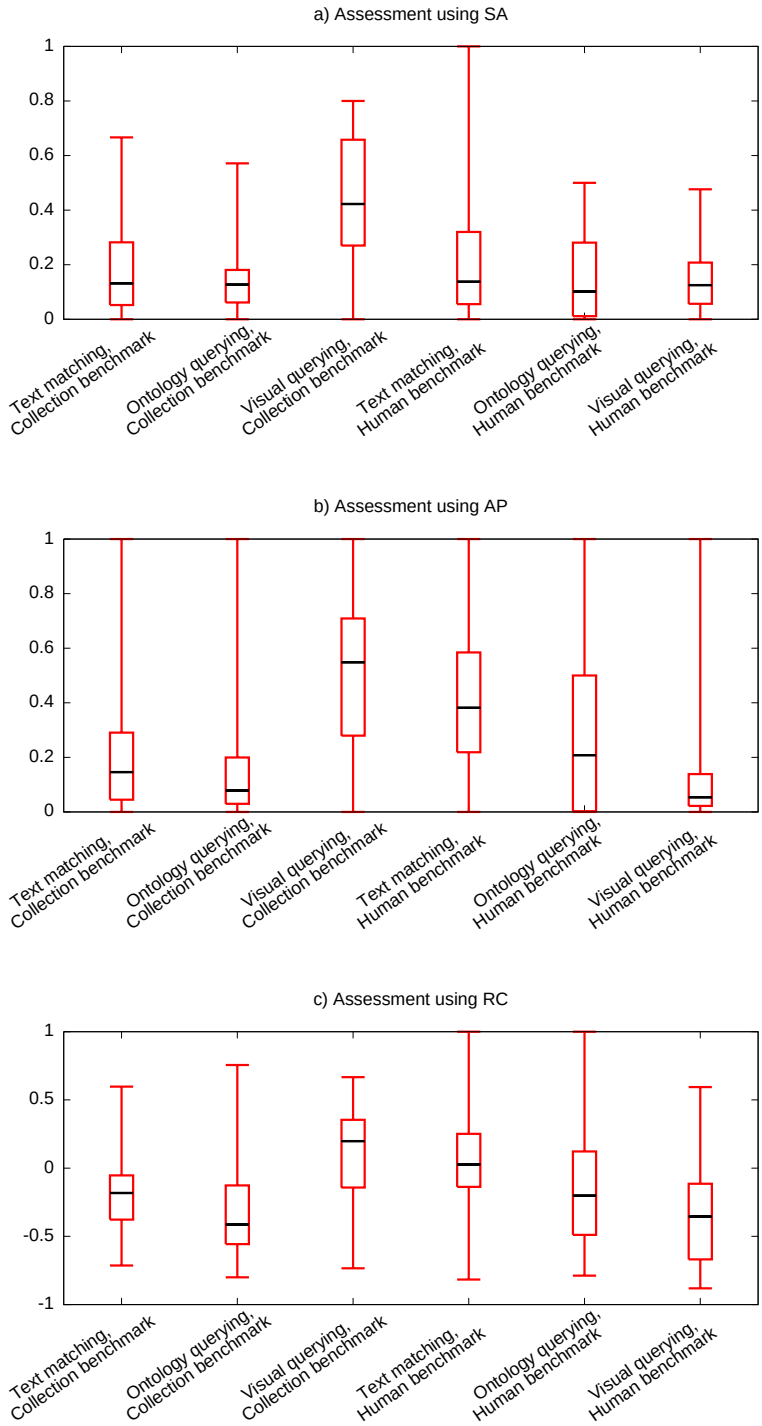


Figure 8.5: Box plots showing the spread of assessment scores for the three automatic concept selection algorithms using the two concept selection benchmarks, over 50 queries. Each assessment measure is given a separate plot.

Table 8.5: Overall accuracy of the concept selection benchmarks in predicting the best performing automatic concept selection algorithm per query, over 50 queries.

	Retrieval prediction accuracy	
Assessment measure	*Collection benchmark*	*Human benchmark*
SA	64%	52%
MAP	58%	36%
RC	50%	34%

Predicting Video Retrieval Performance?

The assessment scores indicate that none of the automatic concept selection algorithm produce query-to-concept mappings that are clearly identified as the best by both benchmarks. This is not surprising, as our analysis of the two benchmarks showed that the query-to-concept mappings they contain can be quite different, to the extent where, for eight of the 50 queries, there is no overlap at all between the query-to-concept mappings of the benchmarks.

We evaluated the retrieval performance of the automatic concept selection algorithms on our test set of shots that are annotated by perfect concept detectors. The best performing automatic concept selection algorithm for retrieval under this setup was the visual querying, which gained a MAP score of 0.049, while the next best performing algorithm was text matching, with a MAP score of 0.037. Ontology querying gained the lowest retrieval performance with a MAP score of 0.016. For the three automatic concept selection algorithms implemented in this chapter, then, the collection benchmark was the best at predicting overall retrieval performance, using any of our proposed assessment measures. But is this also the case when assessing the retrieval performance of individual queries?

We also investigated prediction accuracy when applying the benchmarks to assessing concept selection methods for individual queries. The predictions were computed by determining which automatic concept selection algorithm produced the best concepts for a query using a given concept selection benchmark and assessment measure. If concepts from that concept selection algorithm also produce the best retrieval performance in terms of average precision, the prediction was considered to be accurate. The results are shown in Table 8.5. It is apparent that the collection benchmark is not only the best at predicting which automatic concept selection algorithm will give the best retrieval results over all the queries, it is also best at predicting the best performing retrieval algorithm at the level of individual queries. From this query level analysis differences between individual assessment measures also become apparent; the simple set overlap measure (which takes no ranking infor-

Table 8.6: Assessing the concept selection benchmarks.

	Summary	
Criterion	*Human benchmark*	*Collection benchmark*
Semantics	Captures human world knowledge, shared understanding of query and concept within a group of people. Sensitive to different query interpretations.	Captures other world knowledge, especially domain-specific knowledge. Is sensitive to noise, especially when there are very few sample in the training set.
Retrieval performance	Reasonable performance on ideal collection. Can predict performance of an automatic concept selection algorithm in retrieval task with 53% accuracy. Does not correctly identify the overall best performing algorithm using any of our proposed assessment measures.	Reasonable performance on ideal collection. Can predict performance of an automatic concept selection algorithm in retrieval task with 64% accuracy. Correctly identifies the overall best performing algorithm using all of our proposed assessment measures.
Reproducibility	Repeated group experiment achieved 78% overlap and agreement on best concept on 80% of the queries.	Can be perfectly reproduced with same collection, but is sensitive to the amount of training data available.

mation at all into account) is the best predictor of retrieval performance, both for the collection benchmark with 64% accuracy and for the human benchmark with 52% accuracy. The rank correlation measure, which takes the maximum amount of ranking information into account, proves the worst predictor of retrieval performance, with 50% and 34% respectively.

8.4 Discussion

Let us now summarize our observations and give recommendations for using the benchmarks to assess concept selection methods.

8.4.1 Analyzing the Benchmarks

Table 8.6 summarizes our analysis of the collection and the human benchmarks according to the criteria described in Section 8.2.

Two additional observations may be of interest to potential users of our concept selection benchmark creation methodology. First, when *adding new queries*, one can use the same focus group setup when using the human benchmark methodology; for the collection benchmark judges will need to create new relevance judgments for the training set. Second, when *adding new concepts* to the human benchmark the focus group effort must be repeated with the new set of concepts. For the collection benchmark, one must annotate the video collection with respect to the new concepts (while retaining existing annotations).

8.4.2 Evaluating Concept Selection Algorithms

Next, we give recommendations to assist users in choosing a concept selection benchmark for the assessment of a new concept selection algorithm. In the experiments that we performed in this chapter, the collection benchmark was consistently the best at identifying the best performing automatic concept selection algorithm, and therefore we recommend that new automatic concept selection algorithms be assessed using this benchmark. However, it is important to note that the benchmarks were tested on data from the same domain as the training data that was used to develop the collection benchmark. If a user is performing concept selection in a different domain, then the human benchmark may be a better choice, as this is not tuned to a specific domain.

Our final recommendation concerns the choice of assessment measure. We recommend that set overlap be used, as this measure most consistently predicts the best performing retrieval method on a per-query basis.

8.5 Conclusions

Automatic concept selection, or the identification of visual concepts that represent a user query, is a part of detector-based search. In this chapter we isolated the task of developing resources to assess automatic concept selection algorithms independently of concept detection approach. Based on the literature, we suggested the development of concept selection benchmarks, containing benchmark query-to-concept mappings that can be used to assess new concept selection algorithms. We identified two orthogonal approaches to developing such benchmarks, implemented them for a shared set of queries, and performed an in-depth analysis of each. Both concept selection benchmarks consist of a set of queries, with every query mapped to concepts from our multimedia thesaurus of 450 visual concepts. We also investigated the effectiveness of each of the benchmarks at assessing three automatic concept selection algorithms.

Our first research question CRQ 1 was *What are the differences in terms of semantics of the query-to-concept mappings produced by collection-based versus human-based development of concept selection benchmarks?* We found that the two benchmarks did not produce perfectly overlapping query-to-concept mappings, and thus they did not assign identical semantic interpretations as to the appropriate concepts to map to a query. However, in a query level analysis of the queries with the highest overlap, we found that perfect overlap was only precluded because one benchmark mapped more concepts than the other. We also found that a collection benchmark and a human benchmark can create very different semantic mappings; for eight queries with no overlap whatsoever in the concepts mapped by the two benchmarks.

In our study of randomly selected examples, we found that the expectations of humans sometimes clearly differed in terms of what they expected to see in relevant shots, as compared to what was actually present in relevant shots in the training collection (for example, where users expected *Tropical Settings* to be an appropriate concept for the query *One or more palm trees*, what was actually present in the Broadcast News collection was palm trees shown in the background of urban warfare with *Weapons* and *Shooting*.

Next we asked CRQ 2, *What is the retrieval performance of the query-to-concept mappings contained in the concept selection benchmarks when applied to retrieval in a collection that has been labelled with visual concepts? In other words, how well do the query-to-concept mappings in the benchmarks work for retrieval when we have perfect detectors?*. We found that the two concept selection benchmarks that we created gave approximately equal performance on a test set of shots with perfect detectors. A query-level analysis showed that query-level performance varied a lot, with the retrieval using collection benchmark mappings performing better when the human benchmark mapping introduced concepts that drifted away from the query, and vice versa, the human benchmark performed better when the collection benchmark mapping introduced drifting concepts (in the studied cases where the latter occurred, this was due to a paucity of relevant shots in the training collection with which to generate a mapping).

Turning to CRQ 3, *How consistently can query-to-concept mappings be reproduced for each of the benchmark creation methods that we describe?*, we found that even when using focus groups to arrive at a shared agreement of the concepts to use for a query, agreement was high at 78% when considering the overlap of the query-to-concept mappings of the first focus group with respect to those of the second focus group. We used this asymmetric overlap measure because the first focus group mapped more concepts, on average, than the second. Furthermore, the two groups agreed on the best concept to be used for a query 80% of the time. For the collection benchmark, we compared the query-to-concept mappings at different collection sizes, agreement with the final collection benchmark was halved when using 10% of the training set to generate the benchmark. Based on these observations, and our previous observations of noise in the collection benchmark when there are very few relevant examples for a query, we feel that the collection-based benchmark development will be more reproducible across different collections within the same domain, provided that the training collection used for development contains enough relevant examples for a query. However, when creating a benchmark for assessing concept selection algorithms that need to work across multiple domains, we expect the human benchmark to be more reliable (as it is collection-independent).

Finally, we demonstrated the use of our concept selection benchmarks for automatic concept selection. In answer to CRQ 4, *How successful is each of the two*

concept selection benchmarks at assessing the comparative performance of automatic concept selection algorithms?, we found that the collection benchmark was best at predicting comparative retrieval performance of the three algorithms that we examined. However, it is unclear whether this would still be the case when applying the concept selection algorithms to a new domain; we leave this to future work.

This brings us to a position where we can discuss the overarching research question that directed this chapter,

> **RQ 3** *Given that search by visual concept detector is a valuable method for content-based video retrieval, how can we identify the correct visual concepts to use for a query?*

We found that assessment of automatic concept selection algorithms can be performed independently of concept detection performance in individual systems, by using an annotated collection to generate an ideal mapping of visual concepts to a query, and then assessing the similarity of a candidate set of selected concepts to this benchmark mapping. We also investigated the application of focus group studies to create a comparable benchmark; while the results were promising in terms of the agreement between experiments, this benchmark was less successful at assessing the comparative performance of automatic concept selection algorithms than the collection benchmark. We leave further investigation of the cross-domain effectiveness of the collection benchmark to future work.

In an additional comment, our original aim was to develop a benchmark of *ideal* query-to-concept mappings for retrieval. In our study, for almost 50% of the queries the query-to-concept mappings exhibited differences in retrieval performance between the two benchmarks, with in some cases the collection benchmark performing best, and in some cases the human benchmark performing best. Overall, the two benchmarks exhibited similar retrieval performance. Thus in no way can the query-to-concept mappings of either benchmark be said to be ideal, despite the knowledge-intensive methods used to create them. The question of whether an ideal query-to-concept mapping can be created remains an open one.

In this chapter we found a visual querying algorithm to work best for detector-based search with perfect detectors; in Chapter 10 we will employ a closely related visual querying algorithm in our experiments with detector-based search with real-world detectors.

Chapter 9

Redundancy and the Temporal Mismatch

Having addressed an aspect of detector-based search in the previous chapter, we will move on in this chapter to explore a problem faced when searching on the basis of automatically generated speech transcripts. In Section 3.3.4 we noted that user keyword searches cover a wide range of words, with the majority of query terms occurring only a few times. Transcript-based search allows us to directly perform retrieval on a wide range of words, and potentially to find visual content. This suggests that transcript-based search may be well-suited to the needs of audiovisual broadcast archives. However, transcript-based search is sensitive to the *temporal mismatch*; a lack of alignment between the spoken mention of a word in the audio signal and its appearance in the video signal. Here we will address the problem of the temporal mismatch by modeling *redundancy* phenomena.

Redundancy is the phenomenon that narrative structure is used to repeat information that is important to a video across multiple shots, and in multiple ways [172]. For example, a person appearing in the video signal may also be mentioned by name in the dialog captured in the audio signal. In addition, the person may appear in multiple shots in close temporal proximity, especially if they are a central figure in the narrative. In this way, if we are temporarily distracted from listening to the dialog or looking at the screen, we still have an idea of the semantic content of the video. As we watch a video segment, we gain understanding of its content by integrating different forms of information over time.

Our working assumption is that redundancy across audio and video signals can be used to improve the effectiveness of video retrieval. In this chapter we are espe-

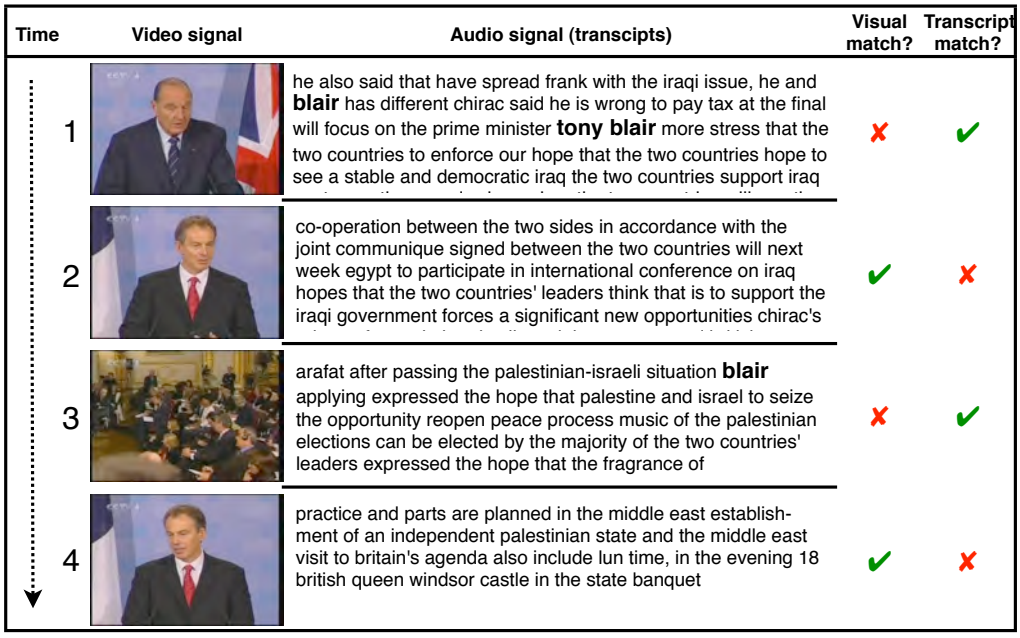


Figure 9.1: Finding Tony Blair: an example of temporal item distribution across the audio and video signals within series of shots from a news broadcast. Tony Blair is seen in the video signal of the second and third shots, while his name appears in transcripts (derived from the audio signal) of the first and third shots.

cially interested in using redundancy—both within the video signal, and across the video and audio signals—to improve the effectiveness of transcript-based retrieval of visual items. To be able to address this issue we first examine the redundancy phenomenon itself. Let us explain. We call a shot *visually relevant* to an object, person, or scene when that particular item can be visually observed in the shot. Now, if a given shot is visually relevant, how likely is it that a neighboring shot is visually relevant as well? And when visual relevance does spread out to neighboring shots, at which distance can we still observe the effect? The phenomenon is illustrated in Figure 9.1. Here we see keyframes extracted from four consecutive shots in a news broadcast featuring Tony Blair, the former prime minister of the United Kingdom. The keyframes contain similar visual and semantic subject matter, with Tony Blair appearing in both shots 2 and 4. How systematic is this phenomenon?

Let us look at another example of redundancy, this time across the audio and visual signals. When objects are difficult to detect on the basis of the video signal, a retrieval engine can look for clues in information captured from the audio signal. For example, individual people such as Tony Blair or Saddam Hussein tend to be

difficult for an automatic system to detect using only visual information. However, a news broadcast showing Tony Blair is likely to mention his name several times, and (depending on the performance of the automatic speech recognition system used) his name will therefore be present in the transcripts. For retrieval, an interesting challenge emerges here: taking into account the *temporal mismatch* that can occur across the audio and the video signals. While each signal is temporally cohesive in its own right, the content may not be synchronized across them. For example, in Figure 9.1 we see a displacement across the mention of Tony Blair's name in shot 1, and his visual appearance in shot 2. It has been shown that named people are on average mentioned two seconds before they appear in broadcast news [195]. We perform an investigation into the distribution of visual to visual relevance (i.e., how likely it is for visual items to occur closely together) and contrast this with the distribution of cross-signal transcript to visual relevance (i.e., how likely it is for items to occur in the visual signal, at given temporal proximity to their mention in the audio signal as captured in transcripts).

Now, assuming that we have developed sufficient understanding of redundancy to model a distribution of the phenomenon, the next step is to try and exploit the phenomenon to improve the effectiveness of transcript-based search for visual items. In a transcript-based search system we could utilise redundancy, and more specifically, the tendency of relevant subject matter to occur close together, by returning temporally related shots at retrieval time. Returning to Figure 9.1, if transcripts indicate that shot 4 is relevant, there is an increased probability that surrounding shots are also relevant. In this study we find consistent redundancy patterns within and across the video and audio signals, and we propose retrieval models that integrate these patterns to improve cross-signal retrieval performance.

We center on the question,

> **RQ 4** *Within a video broadcast, the same object may appear multiple times within and across the video and audio signal, for example being mentioned in speech and then appearing in the visual signal. How can this phenomenon be characterized, and can we model and use this characteristic so as to improve cross-stream retrieval of visual items using transcripts?*

More specific questions are:

CRQ 1 Given a set of visual items in the form of queries or concepts, how are visually relevant shots distributed across shots in the video signal?

CRQ 2 How can we characterize the temporal mismatch for visual items across the video and the audio signal?

CRQ 3 How consistent are our characterizations between collections?

CRQ 4 What is the effect on the performance of transcript-based search of incorporating characterizations of redundancy and the temporal mismatch?

We address our questions through an empirical exploration of real-world video data and by developing a retrieval framework for incorporating temporal redundancy distributions in transcript-based search.

The remainder of the chapter is organised as follows. Section 9.1 is devoted to characterizing the redundancy phenomenon. Section 9.2 outlines the retrieval framework in which we incorporate redundancy models, and in Section 9.3 we describe the retrieval results and analysis. Conclusions are presented in Section 9.4.

9.1 Discovering Redundancy

Here we describe our studies into the characterization of redundancy and the temporal mismatch within and across the audio and video signals in audiovisual material.

9.1.1 Methodology

In order to answer CRQ 1 (about the temporal distribution of visually relevant items in the video signal), we have to characterize the redundancy of a visually relevant video item across time. Our approach in answering this question follows the quantitative approach taken by Yang and Hauptmann [194], who are interested in the *transitional probability* of a neighbouring shot e being visually relevant to an item, given that the previous shot d is visually relevant. We extend the approach to include shots more than one step away from d, in order to allow us to calculate the distribution of transitional probabilities over a shot neighbourhood. The transitional probability is then estimated by $\hat{p}(e_n = V_R|d = V_R)$, where V_R indicates that a shot is visually relevant, and n is the number of shots between e and d. In cases where e occurs before d, n is negative. We use the manually created judgments of V_R to calculate the transitional probability of an item at offset n according to

$$\hat{p}(e_n = V_R|d = V_R) = \frac{c(e_n = V_R, d = V_R)}{c(d = V_R)}, \tag{9.1}$$

where $c(e_n = V_R, d = V_R)$ is the count of the shot pairs where both e_n and d are visually relevant to the item, and $c(d = V_R)$ is the total number of visually relevant shots in the collection. When there is no shot at offset n, due to the offset being outside the beginning or end of the video, then $e_n \neq V_R$.

To answer CRQ 2 (about the temporal mismatch across the audio and visual signals) we calculate the transitional probability of e_n given that d has a match in

Table 9.1: Overview of the collections and visual items used in the experiments in this chapter. Mean numbers of assessments of V_R and V_T per visual item are indicated by $\bar{x}(V_R)$ and $\bar{x}(T_R)$ respectively. Note that for queries $\bar{x}(V_R) < \bar{x}(T_R)$, while for concepts $\bar{x}(V_R) > \bar{x}(T_R)$.

		Item Statistics		
Collection	**Item Type**	*# items*	$\bar{x}(V_R)$	$\bar{x}(T_R)$
Broadcast News	Query	67	47	874
Broadcast News	Concept	450	1,189	201
Broadcast News+	Query	96	203	889
Archive Footage	Query	101	252	933

the *transcript* stream. Substituting the transcript relevance, T_R of d into Eq. 9.1 this gives

$$\hat{p}(e_n = V_R | d = T_R) = \frac{c(e_n = V_R, d = T_R)}{c(d = T_R)}, \tag{9.2}$$

where we say that $d = T_R$ when the transcript associated with the shot contains one of the words of the item description.

9.1.2 Experimental Setup

Collections For our experiments we require a collection of shots that can be associated with shot-level assessments of V_R and T_R for different queries. An overview of the collections in terms of the numbers of queries and assessments of V_R and T_R is given in Table 9.1.

Both of the collections described in Chapter 7 provide us with such data; each collection contains queries and manually created judgments of V_R, and each shot is associated with automatically created transcripts of speech that can be used to assess T_R. In addition to queries, the Broadcast News collection contains judgments for the 450 visual concepts contained in the multimedia thesaurus; we use these concepts as an additional set of visual items.

Not all of the query sets in the Archive Footage collection are suited for characterizing redundancy. Specifically, we exclude the Archive query set from our experiments, as it not associated with manually created judgments of V_R; the judgements for these queries were created using implicit data obtained from user purchases, and therefore video segments consisting of multiple shots, and sometimes entire programs, are grouped together (see Section 7.2.2). Therefore we exclude these from our experiments in this chapter, giving a combined set of 101 queries for the Archive Footage collection.

In order to provide additional data for evaluation and comparison, we define one

extra collection for our experiments, which we term the *Broadcast News+* collection. This collection is derived from the data sets used for TRECVID $2003 - 2006$ benchmarking evaluations. Like the Broadcast News collection, the video data in this collection consists of English, Chinese, and Arabic news broadcasts, and is accompanied by queries, manually created relevance judgments, and (machine translated) English language transcripts. It differs to the Broadcast News collection in that it is not annotated with respect to the multimedia thesaurus of 450 concepts. In total this collection contains over 190,000 shots, and is associated with 96 textual queries as outlined in Table 9.1.

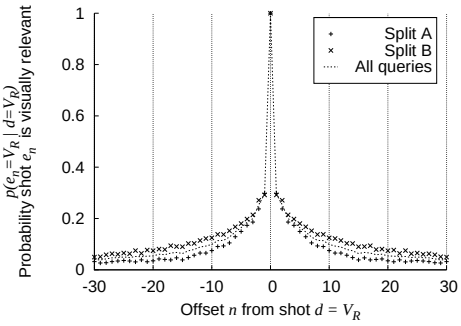
Assessing T_R The transcripts associated with each collection are used to assess T_R for the shots in the video data. This is done by matching text from the queries and concept descriptions to the text in the transcript of a shot. If a term from the transcript of d is matched for a query or concept, then $d = T_R$. Concept descriptions consist of synonyms (synsets) obtained from WordNet using the links from the unified multimedia thesaurus. Query descriptions consist of the natural language descriptions of information need described in Chapter 7.

Partitioning visual items We randomly split the visual items for each combination of collection and item type into two sets, *Split A* and *Split B*. We do this to avoid overfitting our models of redundancy patterns to our retrieval experiments. In our retrieval experiments we will use models developed on Split A when performing retrieval for items in Split B, and vice versa.

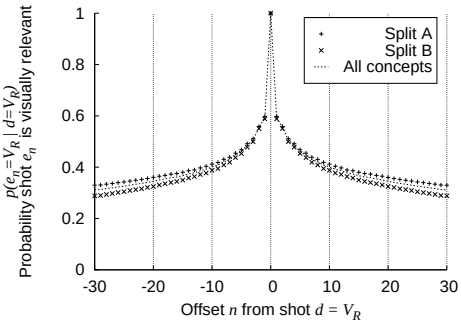
9.1.3 Redundancy in the Video Signal

Given that the current shot contains a visually relevant item, what is the probability that a neighboring shot is also visually relevant? Figure 9.2 gives an overview of the transitional probabilities, calculated using Equation 9.1 and averaged over all instances of $d = V_R$, for the queries and concepts contained in the Broadcast News, Broadcast News+, and Archive Footage collections. The graphs are centered around the known visually relevant shot $d = V_R$ in the middle; along the X-axis we plot the distance from this shot as measured in terms of the number of shots.

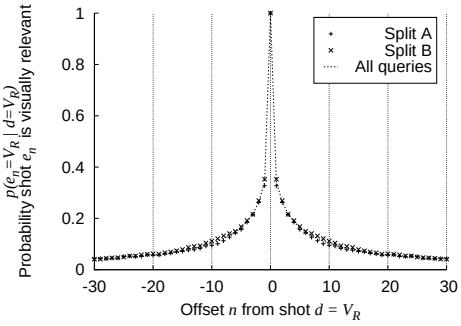
In Figure 9.2 concepts and queries are plotted separately for the different collections, and we see that the redundancy patterns exhibit a similar shape. They are all symmetrical, each pattern peaks sharply at the shot offset of 0 (the known visually relevant shot), and each pattern smooths out to the background probability that any random shot is visually relevant. These curves resemble a power law distribution, as we will discuss in Section 9.1.5. This contrasts to the Gaussian distribution observed by Yang et al. [195]; this because they defined the redundancy window in



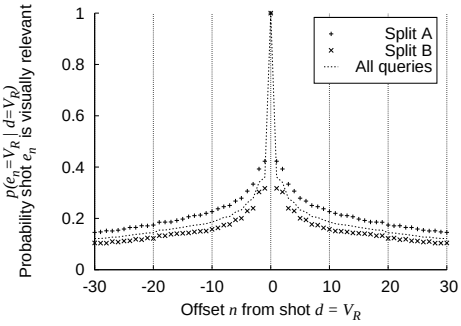
(a) Broadcast News queries



(b) Broadcast News concepts



(c) Broadcast News+ queries

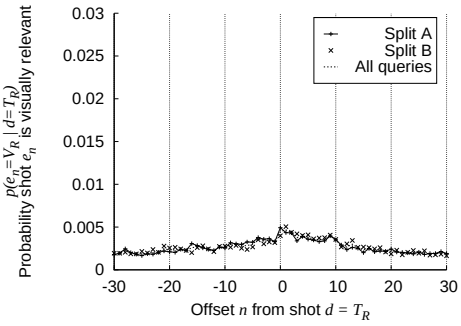


(d) Archive Footage queries

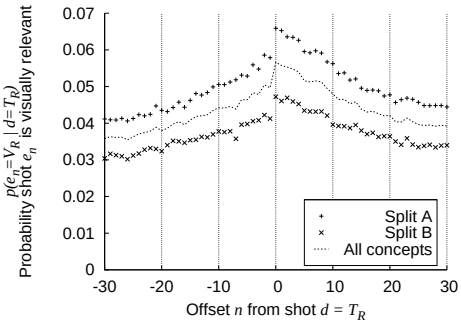
Figure 9.2: Visual redundancy patterns showing the transitional probability of a shot being visually relevant, based on its offset from a known visually relevant shot

terms of seconds, resulting in a smoother distribution. We, on the other hand, to enable a larger scale of analysis, have defined the redundancy windows in terms of shots. This choice was made because plentiful shot-level relevance data is available, however, shots are units that encompass uneven amounts of time, and this makes prevents the measurements from being evenly distributed.

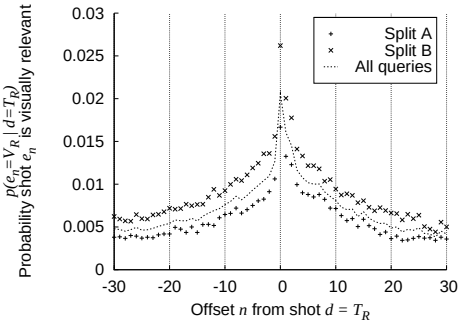
There is also a notable difference between the redundancy patterns, specifically between that for concepts in the Broadcast News collection shown in Figure 9.2b and the query distributions on the different collections shown in Figures 9.2a, 9.2c, and 9.2d. The concept pattern smooths out to a higher probability value than the query patterns do. This is because concepts have more visually relevant shots on average than queries do. For example, the concept with the most relevant shots in the Broadcast News collection, *Person*, occurs in 33,869 of the 43,907 shots, and has a background probability of 0.95. In contrast, the query with the most relevant shots



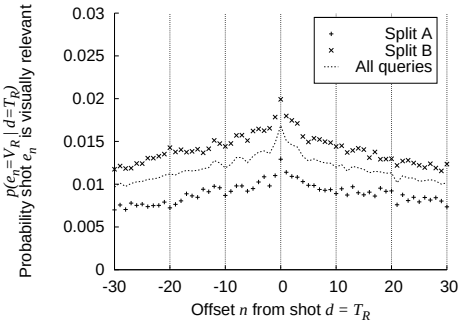
(a) Broadcast News queries



(b) Broadcast News concepts



(c) Broadcast News+ queries



(d) Archive Footage topic

Figure 9.3: Cross-signal redundancy patterns showing the transitional probability of a shot being visually relevant, based on its offset from a match in the transcripts. Note the difference in scale for the Broadcast News concepts, which on average have more relevant shots in the collection than do the queries.

in the Broadcast News collection, *Presidential candidates*, occurs in 240 shots and has a background probability of only 0.01. With concepts occurring more frequently than queries in the shots in the collection, a given concept is more likely to occur in a randomly selected shot than a given query.

9.1.4 Redundancy Across the Video and Audio Signals

Figure 9.3 shows the transitional probability of visual relevance for surrounding shots, given that the transcript of a shot $d = T_R$ contains an item word. Note the scale differences compared to Figure 9.2, which demonstrate that T_R by no means directly indicates V_R; when the transcript of a given shot contains a query or concept

Table 9.2: The average probability that a neighbouring shot is visually relevant, for the 30 shots before the text match and the 30 shots after the text match. The difference in the average probability before and after the text match is indicated by Δ. In collections of broadcast news, visually relevant shots are more likely to occur after a text match than they are to occur before a text match.

		Average probability $e_n = V_R$		
Collection	**Item Type**	$n \in [-5 \,..\, -1]$	$n \in [1 \,..\, 5]$	Δ
Broadcast News	Query	0.0025	0.0028	10%
Broadcast News	Concept	0.0413	0.0454	10%
Broadcast News+	Query	0.0068	0.0072	6%
Archive Footage	Query	0.0118	0.0118	0%

word, the probability that the shot is visually relevant is still very low. Though the scale has changed, we see evidence of redundancy across the signal sources for both queries and concepts: there is a peak in probability of visual relevance close to the point where a transcript match occurs. This is most pronounced for the Broadcast News+ queries.

Furthermore, unlike redundancy within the visual signal, the distributions here are not symmetrical around the central shot d. There is evidence of a temporal mismatch, with shots being more likely to be visually relevant after the transcript match than they are before the match. This is especially apparent in Figure 9.3b, where the curve after $d = T_R$ is less steep than the curve after it. We quantify the temporal mismatch effect over the first 30 shots before and after the text match in Table 9.2. For all the item types in the collections based on broadcast news, a visually relevant shot is more likely to appear after the text match than before the text match. For the queries in the Archive Footage collection, a visually relevant shot is equally likely to appear before the text match is it is to appear afterwards.

We have now uncovered redundancy phenomena and a temporal mismatch across the video and the audio signals in video; next we will model these empirical phenomena, so that we can integrate them into a retrieval system.

9.1.5 Estimating Expected Visual Relevance

This section is devoted to estimating the expected visual relevance e_n for the shots surrounding d. We will estimate γ in two ways: (1) by the empirically derived distribution, and (2) by a simplified approximation of the distribution.

To estimate γ from empirical redundancy patterns, we scale each pattern so that $\gamma = 1$ at the maximum value, and $\gamma = 0$ when the transitional probability is equal to the background probability of the item occurring anywhere in the collection.

To estimate γ by an approximation of the empirical redundancy patterns, we

build on the observation made previously that the visual redundancy patterns resemble power law distributions. We use logistic regression on the data points to develop power law functions of the form $\gamma = bx^m$, where b and m are constant, and x is the absolute offset from the shot with the highest visual relevance probability. In the case of cross-signal redundancy patterns, where the data-points are asymmetrical on either side of the centre, we regress a separate power law function for each side of the curve. This power law will be used to inform the retrieval model that we develop in the next section.

9.2 Retrieval Framework

Our exploration of redundancy has shown that the visual content of a shot is to some extent reflected in the transcript of surrounding shots. Therefore, we wish to adjust the transcript of each shot with transcript text from the surrounding shot neighbourhood, and to examine the effect of incorporating our estimations of visual relevance into retrieval. To do this we must develop a retrieval framework for transcript-based search.

9.2.1 Retrieval Based on Language Modeling

We base our retrieval framework within the language modeling paradigm. We choose language modeling as it is a theoretically transparent retrieval approach and has been shown to be competitive in terms of retrieval effectiveness [60, 122, 197]. Furthermore, the philosophy behind language modeling fits well with our retrieval wishes. Let us explain.

Under the standard language modeling approach, we assume that a document d is generated by a random sample of unigrams from a hidden document model θ_d, where θ_d is a document-specific probability distribution [96]. At retrieval time, for a query q, each document is ranked with respect to the probability that q was generated by θ_d. Therefore, the essential problem is estimating θ_d. Assuming an estimated model $\hat{\theta_d}$ of document d and a query q containing words w_1, w_2, ..., w_m, we rank d according to $p(q|\hat{\theta}_d)$ so that

$$p(q|\hat{\theta}_d) = \sum_{w \in q} p(w|\hat{\theta}_d).$$

One approach to determining $p(w|\hat{\theta}_d)$ is to use maximum likelihood estimation (MLE). A simple MLE estimate of $p(w|\hat{\theta}_d)$ is given by $\frac{c(w,d)}{|d|}$ where $c(w,d)$ is the count of w in d and $|d|$ is the total number of words in the document. However, the MLE assigns no probability mass to unseen words, and in addition does not take into account background probabilities of words that occur frequently in the overall doc-

ument collection. Therefore, some type of *smoothing* is commonly used to adjust for (at least) these factors. In our experiments we use the Jelinek-Mercer smoothing method [197], as we have previously found this to be a suited method for transcript-based video retrieval [71]. This method interpolates the maximum likelihood with the background collection language model θ_C. The Jelinek-Mercer smoothing estimate is given by

$$p(w|\hat{\theta}_d) = \lambda \cdot \frac{c(w,d)}{|d|} + (1-\lambda) \cdot p(w|\theta_C), \tag{9.3}$$

where λ is a fixed parameter that controls the interpolation.

Redundancy will be integrated within our retrieval model by adjusting the word counts $c(w,d)$, as we will now explain. A document is not necessarily a complete reflection of its underlying model, and we can use external information to help estimate θ_d. In our approach, documents are shots associated with transcripts. We use redundancy information to help estimate θ_d, and adjust the word counts for a document with transcripts from surrounding shots to help estimate θ_d for each shot.

9.2.2 Document Expansion

Document expansion is a technique originating from spoken document retrieval that allows for incorporation of external evidence in a natural way [140]. In this approach, a document is expanded and re-weighted with related text at indexing time. Traditionally, this approach is used to augment the original document with text from multiple related documents that have been obtained by some form of feedback. In our approach, document 'relatedness' will be assigned according to temporal proximity. Tao et al. [164] propose a general model for document expansion in the language modelling setting, on which we build. To perform document expansion, we use a set of external shots E to determine additional information about every shot d in a collection C. At indexing time we use word counts from the transcripts associated with d and from transcripts associated with the shots in E to create a transcript associated with a 'pseudo-shot,' d'. The word counts in d', $c(w,d')$, are adjusted from those in d according to:

$$c(w,d') = \alpha \cdot c(w,d) + (1-\alpha) \cdot \sum_{e \in E} (\gamma_d(e) \cdot c(w,e)), \tag{9.4}$$

where α is a constant, e is a shot in E, γ is our confidence that e provides information that is useful for d, and $c(w,d)$ is the number of occurrences of w in the transcript of d.

Placing this model in the context of temporally related video data, we have the following:

- our central shot d, and its associated transcript;
- our set of external shots E, which we define as the neighbouring shots within a window of n shots;
- γ is our model of the expected visual relevance for $e \in E$.
- as d is always the central member of E, this eliminates the need for α, which is replaced by the γ value at offset 0.

This leads to the following temporally expanded document model:

$$c(w, d') = \sum_{e \in E} (\gamma_d(e) \cdot c(w, e)), \tag{9.5}$$

We arrive at different retrieval models by making different choices for γ; then, Eq. 9.5 is used instead of the original word count $c(w, d)$ in Eq. 9.3.

9.2.3 Integrating Redundancy

Finally, we need to put together the two main ingredients developed so far: our retrieval framework and the models of expected visual relevance described in Section 9.1.5.

We integrate redundancy into our framework by modifying the γ function in Eq. 9.5. We consider two baselines; for the first no document expansion is performed at all, and for the second document expansion is without incorporating any information about the expected visual relevance of neighbouring shots. We develop four retrieval models based on the different models of expected visual relevance described in Section 9.1.5. The retrieval experiments are denoted as follows:

B1. No expansion. No integrated redundancy; use the model described in Eq. 9.3;

B2. Flat. $\gamma = 1$: all shots are expected to be equally visually relevant;

D1. Visual data driven. γ is determined by the empirical visual redundancy value, at distance n;

D2. Transcript data driven. γ is determined by the empirical cross-signal redundancy value, at distance n;

P1. Visual model driven. γ is determined by a power law approximation of visual redundancy, at distance n.

P2. Transcript model driven. γ is determined by a power law approximation of cross-signal redundancy, at distance n;

Table 9.3: MAP scores for the different retrieval models at window size 30. The highest performing score for each combination fo item type and collection is indicated in bold. ▲, ▼, and ◦, respectively indicate that a score is significantly better than, worse than, or statistically indistinguishable from the scores of B1 and B2, from left to right.

		Retrieval model					
		Baselines		*Data driven*		*Power law*	
Collection	**Item type**	*B1*	*B2*	*D1*	*D2*	*P1*	*P2*
Broadcast News	Query	0.006	0.005	0.010▲▲	0.008◦▲	**0.011**▲▲	0.008◦▲
Broadcast News	Concept	0.005	0.011	**0.013**▲▲	0.012▲▲	**0.013**▲▲	**0.013**▲▲
Broadcast News+	Query	0.040	0.030	**0.076**▲▲	0.069▲▲	0.073▲▲	0.069▲▲
Archive Footage	Query	0.006	0.021	0.022▲◦	0.023▲◦	0.020▲◦	**0.025**▲◦

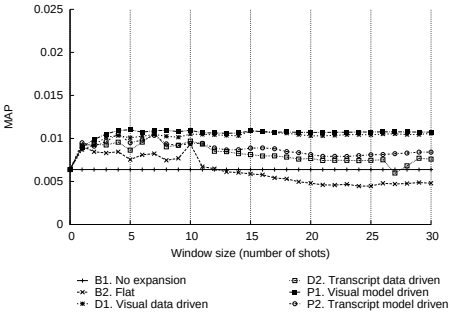
9.3 Retrieval Experiments

In our retrieval experiments we use the retrieval framework developed in Section 9.2 to address CRQ 4, *What is the effect on the performance of transcript-based search of incorporating characterizations of redundancy and the temporal mismatch?* We test each of the retrieval models described in Section 9.2.3 using the data sets and items described in Section 9.1. The models are evaluated at increasing shot window sizes.

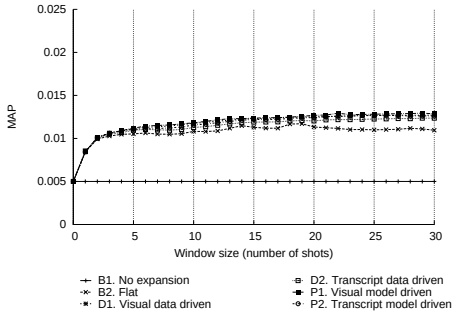
9.3.1 Result Overview

Figure 9.4 provides an overview of the retrieval results for each of the retrieval models that we consider across increasing shot window sizes, while Table 9.3 specifies in numbers the retrieval scores of the retrieval models, and significant differences to the baseline, at a window size of 30 shots. In general we observe that the MAP scores initially increase as the window size increases, indicating that using document expansion always increases overall performance when retrieving visual items using transcripts. However, as transcripts from increasing numbers of shots are incorporated in the expanded documents, differences between retrieval models become clear, as we will discuss for each combination of collection and item type in turn.

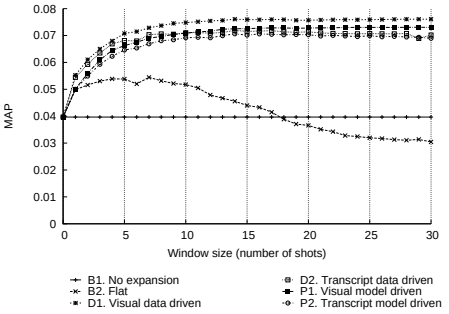
First, let us examine the performance of the retrieval models on the queries in the Broadcast News collection, shown in Figure 9.4a. Here the two models based on visual redundancy, D1 and P1, attain the highest performance, leveling off at a window size of between 5–10 shots. The two models based on cross-signal redundancy patterns, D2 and P2, exhibit different behavior — rather than leveling off as D1 and P1 do, performance starts to degrade at a window size of 5 shots. D2 and P2 do not perform significantly better than B1 at a window size of 30 shots, though they do perform significantly better than B2, which also degrades as more shots are added to the expanded documents.



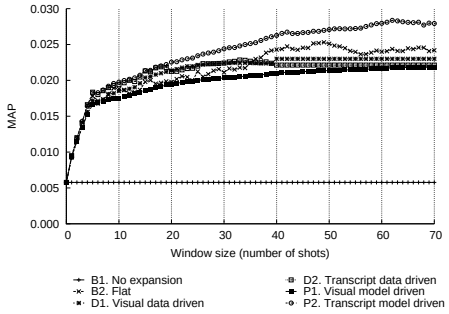
(a) Query MAP scores on the Broadcast News collection



(b) Concept MAP scores on the Broadcast News collection



(c) Query MAP scores on the Broadcast News+ collection



(d) Query MAP scores on the Archive Footage collection

Figure 9.4: Results of retrieval experiments across increasing window sizes. At a window size of 30, the transcript of each shot is expanded with the text from the transcripts of the 30 preceding shots and the 30 subsequent shots. The retrieval performance of the queries on the Audiovisual Archive collection keeps rising to a window size of 62 shots.

The performance of the retrieval models on the concepts in the Broadcast News collection exhibits a different pattern, as we can see in Figure 9.4b. Here all of the redundancy based retrieval models have similar retrieval performance as window size increases, leveling off at a window size of about 20 shots. Significance tests at a window size of 30 shots showed that there is no significant difference between the scores of D1, D2, P1, and P2. B2 similarly increases in retrieval performance, but reaches a lower maximum than the four retrieval models based on redundancy patterns, which significantly improve upon it.

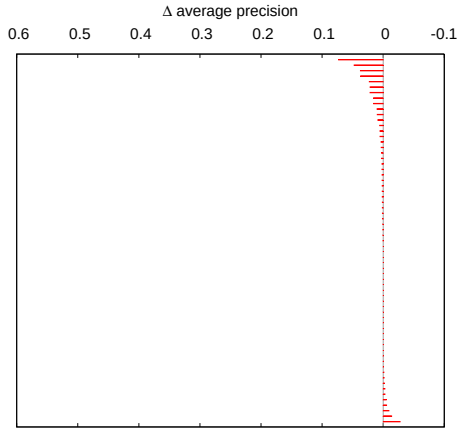
Differences between retrieval models based on redundancy and the two baselines are the greatest, in terms of absolute MAP scores, for the queries in the Broadcast News+ collection. Absolute retrieval performance here is so much higher than that of

the remaining visual items and collections that, for visualization purposes, the scale of Figure 9.4c had to be altered as compared to the remaining figures. Once again, the performance of all four models based on redundancy patterns increases steadily as window size increases, leveling off at a window size of about 15 shots. The two best performing models, D1 and P1, are based on visual redundancy patterns. Of the two, D1, which is based on empirical data measurements, outperforms P2, which is based on power law approximations of those measurements (though not significantly so). Baseline B2 starts to taper off rapidly when using a window size of more than 7 shots, and D1, D2, P1, and P2 all significantly outperform both baselines at a window size of 30 shots.

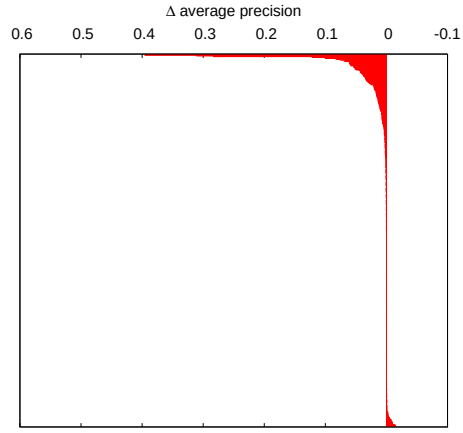
Finally, we turn to the performance of our retrieval models on queries from the Archive Footage collection, shown in Figure 9.4d. The performance here is unusual in that it is still increasing for all models at the maximum shown window size of 30. For this collection we performed retrieval experiments with windows of up to 70 shots, and found that retrieval performance levels out at a distance of 62 shots, with a MAP score of 0.028 for the best performing retrieval model, P2. At this point P2 is significantly better than both baselines. As for baseline B2, this also improves steadily as window size increases, though its performance is significantly worse than P2 at window size 30.

9.3.2 Impact at the Item Level

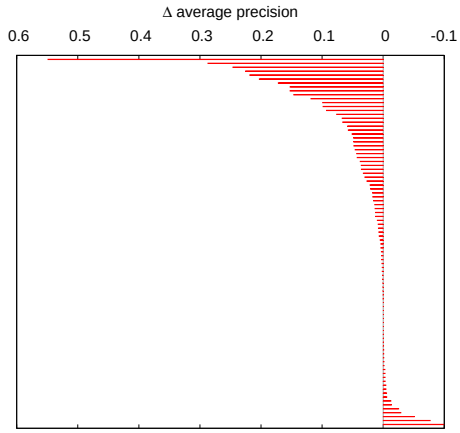
We turn to a discussion of responsiveness of individual items in our collections to redundancy-based models. Figure 9.5 gives an item-level specification of the change in average precision when comparing results from the best performing retrieval model to results from our first baseline, B1, which uses no document expansion at all, at window size 30. From this figure we can see that the majority of visual items benefit from incorporating document expansion. Taking into consideration all visual items across all collections, and taking into consideration changes of magnitude > 0.01 only, retrieval performance is improved for 21% of the visual items, while it degrades for only 2% of the visual items. The largest absolute changes in performance can be observed for the queries in the Broadcast News+ collection. The query that improves the most ($\Delta = 0.549$) is *Find shots of the Sphinx*. A manual inspection of the results revealed that there is one news story discussing the Sphinx in the collection, and it shows 12 visually relevant shots within a window of 15. At the same time the word "Sphinx" occurs only three times in the transcripts of the shots in this story, so by expanding the transcripts the relevant shots are retrieved and performance increases. The query that decreases most in performance ($\Delta = -0.104$) is *Find shots of the Tomb of the Unknown Soldier at Arlington National Cemetery*. A manual inspection of these results showed that the 31 visually relevant shots for



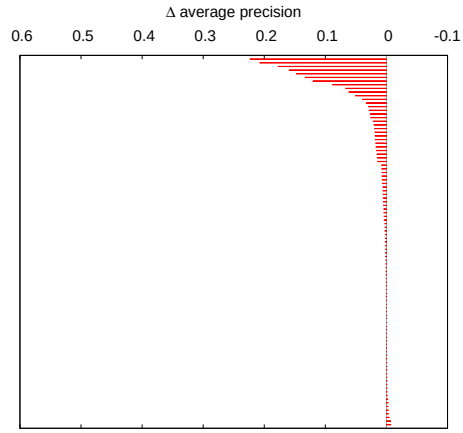
(a) 67 queries, Broadcast News collection



(b) 450 concepts, Broadcast News collection



(c) 96 queries, Broadcast News+ collection



(d) 101 queries, Archive Footage collection

Figure 9.5: Increase Δ in average precision when using the optimal retrieval model at window size = 30 as compared to baseline B1, for the visual items in the Broadcast News, Broadcast News+, and Archive Footage collections. The visual items on the Y axis are sorted by increase over the baseline.

this query are distributed across six different news stories in the collection, and appear in tight clusters. Of the visually relevant shots, 17 contain at least one of the query words and are given a high ranking before expansion. Document expansion introduces irrelevant shots, decreasing the ranking of the relevant results.

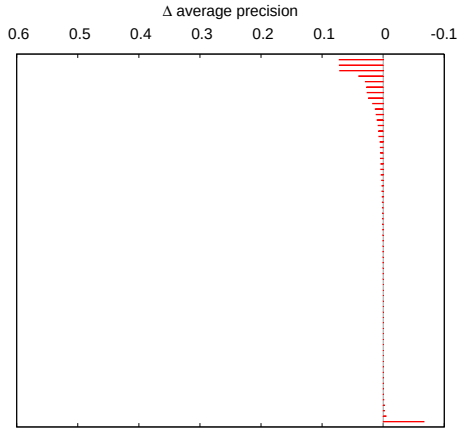
Figure 9.6 gives a specification of the change in average precision when comparing results from the best performing retrieval model to the second baseline, B2. This allows us to compare our redundancy models to a model that does include document expansion, but that does not take redundancy patterns into account. Once again, the majority of items benefit from a retrieval model that includes redundancy patterns. A total of 14% of the items are positively affected, and 3% of the items are negatively effected. The largest increase in performance ($\Delta = 0.398$) is observed for a query from the Broadcast News collection, *Find shots of Boris Yeltsin*. The largest decrease in performance ($\Delta = -0.094$ is for a query from the Archive Footage collection, *Find shots of apes or monkeys*. This query performs poorly as compared to flat expansion because most of the relevant shots are contained in a single documentary program about zoos. These shots are are distributed throughout the program as background footage, while monkeys are only mentioned a few times. Therefore a flat window approach performs better here.

9.3.3 Comparing the redundancy models

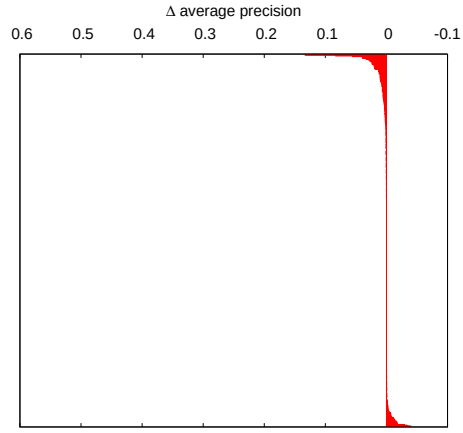
From our experiments it is clear that transcript-based search benefits in all cases from incorporating a retrieval model that takes redundancy into account. Now we turn to an examination of the differences across the individual retrieval models based on the redundancy patterns.

Visual redundancy vs cross-signal redundancy First we discuss the differences between the performance of retrieval models based on visual redundancy patterns (D1 and P1) as compared to the performance of retrieval models based on cross-signal redundancy patterns.

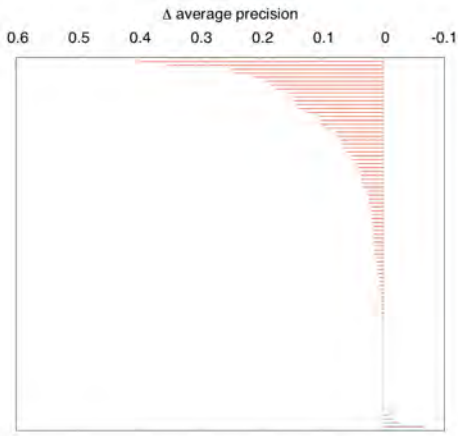
In two of the four cases, a redundancy model based on visual redundancy outperformed a model based on cross-signal redundancy. We found this surprising, as the cross-signal redundancy patterns assign shots occurring after the mention of a word in the transcript a higher estimation of visual relevance than shots that occur before the mention. Models using visual redundancy patterns, on the other hand, assign estimations of visual relevance symmetrically around d. So why did the queries in the Broadcast News and Broadcast News+ collections not respond as well to the models based on cross-signal redundancy? Turning back to the visual representations of the redundancy patterns in Figure 9.3d, we see that for the Broadcast News queries the visual redundancy patterns between splits do not agree. For Split A shot d has the maximum probability of being visually relevant, while for Split B shot $d+1$ has the maximum probability of being visually relevant. For retrieval the splits are switched, causing incorrect estimations of visual relevance using the cross-signal redundancy models. As for the queries in the Broadcast News+ collection, we can



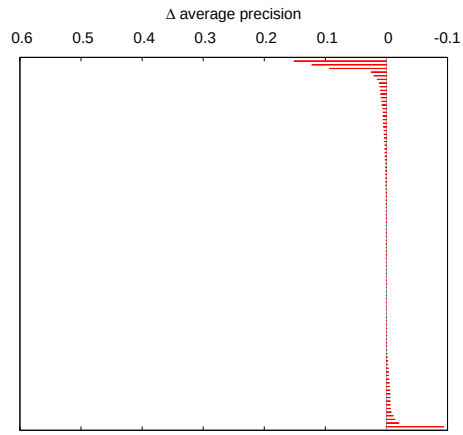
(a) 67 queries, Broadcast News collection



(b) 450 concepts, Broadcast News collection



(c) 96 queries, Broadcast News+ collection



(d) 101 queries, Archive Footage collection

Figure 9.6: Increase Δ in average precision when using the optimal retrieval model at window size = 30 as compared to baseline B2, for the visual items in the Broadcast News, Broadcast News+, and Archive Footage collections. The visual items on the Y axis are sorted by increase over the baseline.

see from Table 9.2 that there is no strong temporal mismatch phenomenon for these queries in this collection, with a visually relevant shot being only 1% more likely to occur in the five shots after d than the five shots before d.

Empirical measurements vs power law approximations Turning to the differences between redundancy models that use empirical measurements (D1 and D2) and redundancy models that use power law approximations (P1 and P2), there is no significant difference between the two kinds of model. Therefore power law approximations may be used to stand in for empirical measurements when applying redundancy models in our framework.

9.4 Conclusions and Future Work

In this chapter we studied the redundancy phenomenon in video retrieval, in a broadcast news setting and audiovisual archive footage, i.e., the phenomenon that in the narrative of video important information is repeated, both within the video signal, and across the audio and video signals. We formulated four chapter level research questions to direct our study.

In answer to CRQ 1, *Given a set of visual items in the form of queries or concepts, how are visually relevant shots distributed across shots in the video signal?*, we found that visual redundancy patterns across shots resemble a power law function; given that a shot is visually relevant to an item, the probability that a neighbouring shot is also visually relevant decreases rapidly as the distance between the two increases. Turning to CRQ 2, *How can we characterize the temporal mismatch for visual items across the video and the audio signal?*, we observed there is redundancy, in that when an item is mentioned in the transcript (derived from the audio signal) of a shot, it is more likely to appear in either that shot or a neighbouring shot than it is to appear in a distant shot. Furthermore, we observed a temporal mismatch, with an item more likely to appear immediately after it is mentioned in the transcript than it is to appear before the transcript. In answer to CRQ 3, *How consistent are our characterizations between collections?*, we measured this phenomenon across different collections, and found that the temporal mismatch effect was was stronger in news broadcasts than it was in more heterogeneous video data obtained from archive footage.

Finally, in order to answer CRQ 4 *What is the effect on the performance of transcript-based search of incorporating characterizations of redundancy and the temporal mismatch?*, we developed a retrieval framework that allows us to incorporate redundancy phenomena in transcript-based search. In this framework, the transcript of a shot is expanded with the (weighted) text of transcripts from surrounding shots. We found that combining transcripts from surrounding shots improved overall retrieval performance, as compared to retrieval performance when search on text from only a single shot, for all our collections and item types. In addition we found that, for all of our collections and item types, retrieval models that expanded transcripts on

the basis of redundancy phenomena outperformed retrieval models that expanded transcripts without incorporating redundancy phenomena. For three out of four collections and item types the increases were significant.

In this chapter we observed that retrieval performance in terms of absolute scores was low. While transcripts can give us some indication of the presence of visual items in the video signal, they are not always effective. In the next chapter we place transcript-based search in the context of the archive of the future, which we postulate will include detector-based search, feature-based search, and transcript-based search, as well as search on the manually created annotations that are already present in the archive.

Chapter 10

Integrating Content-Based Video Retrieval into the Archive

Now that we have examined concept selection methods for detector-based search in Chapter 8, and temporal redundancy between transcripts and visual items in Chapter 9, we move on to examine their potential to improve search in the audiovisual broadcast archive.

Our central aim in this chapter is to answer the final research question,

> **RQ 5** *What is the potential impact of content-based video retrieval in the audiovisual broadcast archive, taking into account both the needs of professional users, and the manually created data already present in the archive?*

As we saw in Chapter 6, existing evaluation initiatives for content-based video retrieval have, in general, relied on queries that are not based on real-world searches. In addition, they have explicitly precluded the use of manually created metadata, which is often present in real world archives. For these reasons, existing evaluation initiatives are not well-suited for investigating content-based video retrieval in the real-world setting of the audiovisual archive. In Chapter 7 we presented the Archive Footage collection, which specifically addresses these issues. Therefore, to achieve our research aim, we propose an evaluation methodology that is based on this collection, and is thereby tailored to the specific needs and circumstances of the audiovisual archive.

In this chapter we answer the following research questions:

CRQ 1 What is the performance of content-based video retrieval when answering

today's queries in the archive, and queries as they might be formulated in the archive of the future?

CRQ 2 What can content-based video retrieval add to search performance when combined with current archive search capabilities?

CRQ 3 Can content-based video retrieval help those users that wish to retrieve not just shots, but entire programs?

CRQ 4 Which content-based video retrieval methods should be given priority for integration into the archive?

Ultimately, our answers to these questions can benefit policy makers at audiovisual archives who are facing the limitations of today's manual annotation practices and are considering incorporating content retrieval into their work-flow. In addition, our answers are of interest to researchers as they apply content-based video retrieval outside of the usual laboratory benchmark setting.

Our evaluation methodology integrates multiple perspectives in order to answer the research questions. First, we employ the three query sets defined in Chapter 7, which include both current user queries and those that might be issued in a future archive, equipped with content-based video retrieval. Second, we create a large set of simulated purchase-query pairs using the best simulator described in Chapter 4, in order to perform large-scale evaluation. Third, we build three video search engines that exploit both manually created and automatically generated annotations. Fourth, we perform and evaluate retrieval at both the shot and program levels. The outcomes of the experiments allow us to explore different ways in which content-based video retrieval might be integrated into tomorrow's audiovisual archive.

The contributions of this quantitative study are four-fold:

- We evaluate the performance of content-based video retrieval for textual queries which are taken directly from the transaction logs of an audiovisual broadcast archive.
- We evaluate the performance of content-based video retrieval for multimedia queries that are developed by reformulating the information needs observed in session-level analysis of the searcher behavior recorded in the transaction logs of an audiovisual broadcast archive.
- We examine the gains that may be achieved by combining content-based video retrieval with retrieval using manually created catalog annotations maintained by the archive.
- We present a publicly available evaluation collection that includes manually created program annotations from the archive, queries based on the information needs of users from the audiovisual archive, and their associated relevance

judgments.[1]

The rest of this chapter is structured as follows. We present our evaluation methodology in Section 10.1. In Section 10.2 we outline our experimental setup. Results are presented in Section 10.3. We end this chapter with conclusions and recommendations for the archive in Section 10.4.

10.1 Evaluation Methodology

We use a quantitative system evaluation methodology to explore the potential of content-based video retrieval for enhancing search performance in the audiovisual archive. Typically, evaluation of retrieval systems requires a collection of documents, a set of statements of information need (called "queries" in this chapter), and relevance judgments indicating which documents in the collection should be returned for each query [176]. Existing evaluation initiatives utilize documents, queries, and relevance judgments that do not reflect retrieval practice in the archive. To remedy this, based on the Archive Footage collection, we incorporate in our methodology: (1) real-world queries derived from archive usage data, as well as laboratory queries used in benchmark evaluations; (2) video search engines based on manually created annotations from the archive; and (3) a program-level as well as a shot-level retrieval task. We summarize our methodology in Figure 10.1 and detail the individual ingredients in the following sections.

10.1.1 Query Definitions

In Chapter 7 we described the definition of three query sets. These three query sets — the Archive query set, the Lab query set, and the Future query set — form the basis of our experimental methodology. The Archive query set represents the queries of searchers as they are issued today, using the catalog-based search engine that is currently available in the archive. The Lab query set represents queries that have been developed by benchmark designers for collections obtained form the audiovisual broadcast archive. The Future query set represents queries as they might be posed by media professionals to the hypothetical search engine of the archive of tomorrow, a search engine that is enabled with content-based video retrieval techniques.

In addition to these query sets, we generate a set of simulated queries using the simulation framework described in Chapter 4. Recall that in our simulation approach, a given document is used to generate a simulated query. The document is then considered relevant to that query. Using a simulator to create a set of queries for evaluation gives us the advantage of being able to create as many queries as we

[1] http://ilps.science.uva.nl/resources/avarchive

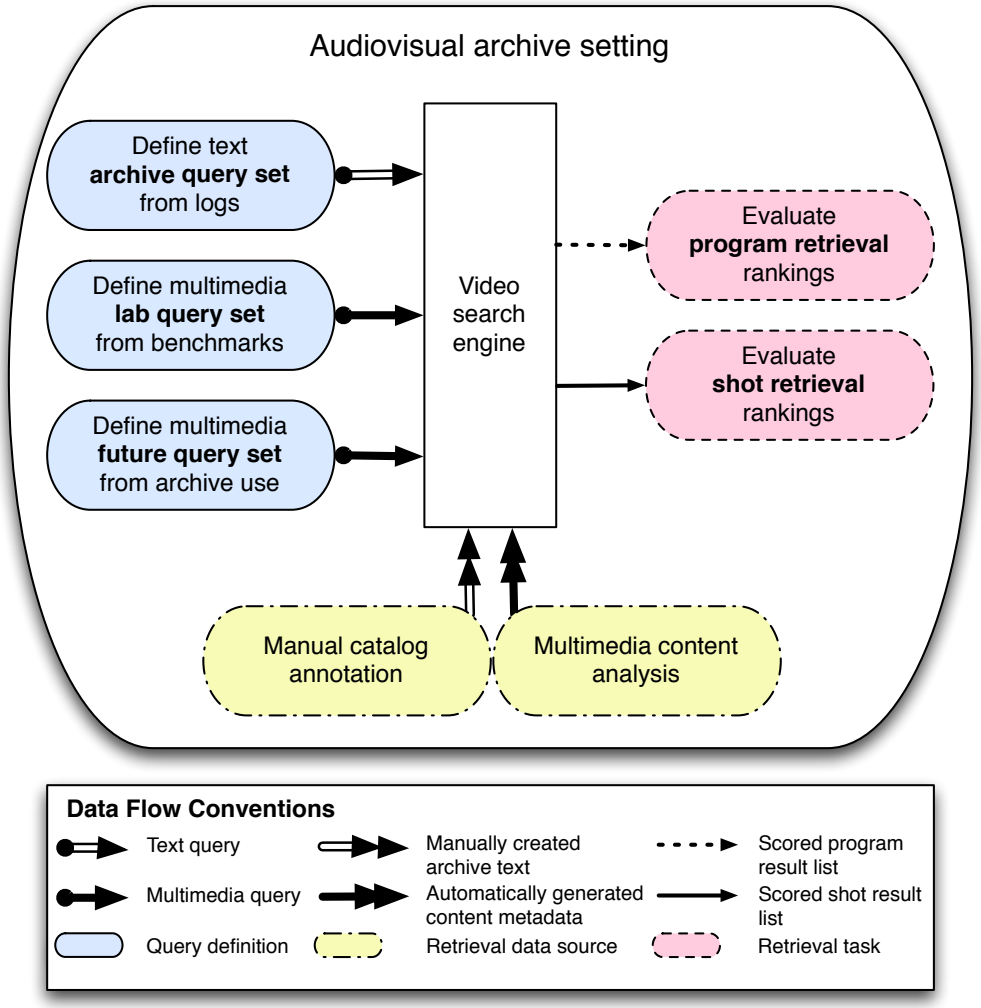


Figure 10.1: Methodology used to evaluate the potential impact of content-based video retrieval in the audiovisual archive.

wish. However there are limitations to this approach. Namely, our simulators create relevance judgments at the level of an entire program, and are therefore not suitable for evaluating shot-level retrieval. In addition, the simulated queries do not necessarily reflect the needs of real users. Keeping these limitations in mind, we generate 10 simulated queries for each of the 219 programs in the Archive Footage collection, resulting in a set of 2,190 simulated purchase-query pairs. For the experiments in this chapter, we use the simulator that that was found to be best in Chapter 4.

10.1.2 Retrieval Data Sources

The retrieval data sources in our evaluation methodology consist not only of automatically generated metadata generated by multimedia content analysis, but also of manually created archive text produced by professional archivists, as we described in Chapter 7. The catalog entries consist of technical metadata, free text descriptions, and tags in the form of person names, video subjects, locations, and so on. The automatically generated metadata consists of current state-of-the-art multimedia analysis results produced by *transcript-based*, *feature-based*, and *detector-based* methods.

10.1.3 Video Retrieval Tasks

We consider two video retrieval tasks.

Shot retrieval Users in the archive cannot currently retrieve shots, but over 66% of the orders in the archive contain requests for video fragments. Shot-based video retrieval could allow these users to search through tomorrow's archive more efficiently. Therefore, we include a shot retrieval task in our evaluation methodology.

Program retrieval Users in the archive currently retrieve entire programs, and tomorrow's archive is likely to continue to support this task. Therefore, we include a program retrieval task in our evaluation methodology. This requires an adaptation of the relevance judgements in our collection, which are at the shot-level. We create relevance judgments at the program level using a simple rule: if a program contains a shot that is relevant to the query, then we consider the entire program relevant to the query.

10.1.4 Video Search Engines

Video search engine 1: catalog-based Our *Catalog* search engine indexes the catalog entries associated with the programs in the collection. The (Dutch language) free text, tags, and technical metadata are each indexed and retrieved separately. We normalize, stem, and decompound [109] the text. Retrieval is done using the language modeling paradigm [122], with the Lemur toolkit implementation. To compensate for data sparseness and zero probability issues, we interpolate document and collection statistics using Jelinek-Mercer smoothing [197]. In addition, as the collection of 219 catalog entries ("programs") provides a relatively small sample from which to estimate collection statistics, we augment these with collection statistics from a sample of 50,000 catalog entries randomly selected from the archive. The

Catalog search engine is based on program-level information; to return results at the shot-level, we return the shots for a program in order of appearance. This reflects the current way in which the archive's interface currently presents keyframe information (see Figure 3.1d on page 25).

Video search engine 2: content-based The *Content* search engine is based on shot-level multimedia content analysis, covering transcript-based, feature-based, and detector-based search. We create a retrieval result for each of the three types of search using the state-of-the-art methods described in Chapter 6, with implementation of Snoek et al. [159]. Since both the detector- and feature-based retrieval methods rely on multimedia query examples as input, we rely on transcript retrieval for the archive-based text-only queries (without multimedia examples). The Content search engine is based on shot-level information; to return results at the program level, we turn to the literature from text retrieval, where *passage based retrieval* has addressed the problem of aggregating results from parts of a document to retrieve entire documents [133]. The methods developed here are directly transferable to our problem, and we use the decay-based method of Wilkinson [186] to aggregate results from the shot level to retrieve entire programs.

Video search engine 3: future The *Future* search engine is formed by selecting the optimal combination of retrieval results from both the catalog- and content-based video search engines. The optimal combination is produced using the result fusion method described in the next paragraph. The merging of search engines reflects a realistic retrieval scenario for the archive of tomorrow, where the manual annotations from the archive have been merged with automatic multimedia content analysis. The engine can be adjusted for program or shot retrieval by varying the unit of the input results.

Result fusion

All three video search engines use multiple lists of search results that we need to combine into a single list for evaluation. To produce this single list we perform fusion using the settings recommended by Wilkins [183] as described in Chapter 6, i.e., we truncate each retrieval result to contain no more than 5,000 items, we normalize the scores using Borda rank-based normalization, and we fuse all results using the weighted CombSUM method. Since we are concerned with evaluating the *potential* of video retrieval in the archive, we simply take for each query the combination that optimizes retrieval performance.

10.2 Experimental Setup

Now that we have outlined our evaluation methodology, we move on to describe the experimental setup.

To answer our research questions related to the potential of content retrieval for improving search performance in the audiovisual broadcast archive, we conduct the following four experiments:

Experiment 1 *Shot retrieval with three video search engines using three query sets.* In this experiment, we address the task of retrieving visually coherent fragments from the archive, a type of search currently unavailable in the archive. We retrieve video fragments using three query sets also, and again with three different video search engines. This experiment aims at answering CRQ 1 and CRQ 2.

Experiment 2 *Program retrieval with three video search engines using three query sets.* In this experiment we emulate the current retrieval practice in the audiovisual archive. We retrieve videos as complete productions using three query sets and with three different video search engines. This experiment aims at answering CRQ 1, CRQ 2 and CRQ 3.

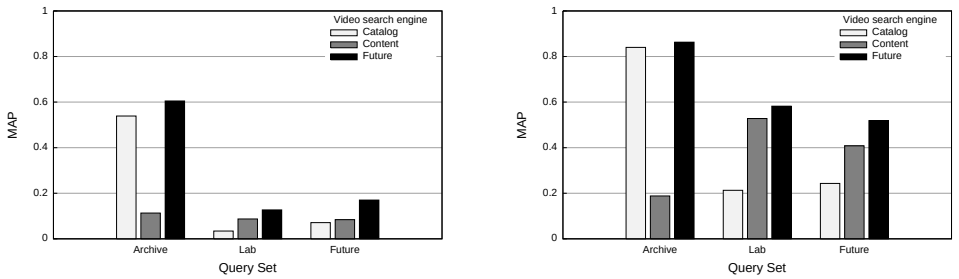
Experiment 3 *Program retrieval with three video search engines using simulated queries.* We create a set of 2,190 simulated purchase-query pairs using the best simulator from Chapter 4, and use these to evaluate program-level retrieval performance of the three different video search engines on a large scale. This experiment is in aid of answering CRQ 1, CRQ 2, and CRQ3.

Experiment 4 *Prioritizing content-based video search methods.* We examine the potential contribution of three types of content-based search: transcript-based search, feature-based search, and detector-based search. This experiment aims at answering CRQ 4. We perform this experiment on the query sets that contain multimedia queries, namely the Lab query set and the Future query set, as the text-only Archive queries cannot be used for all three types of search.

Performance measure and significance tests As detailed in Chapter 6, for all four experiments, we evaluate the top 1,000 ranked shot- or program-level results using the standard mean average precision (MAP) measure. In addition, we perform Wilcoxon Signed Rank tests at the 0.01 level for significance tests.

	Experiment 1: 3x3 Shot Retrieval			Experiment 2: 3x3 Program Retrieval		
	Video search engine			*Video search engine*		
Query set	*Catalog*	*Content*	*Future*	*Catalog*	*Content*	*Future*
Archive	0.539	0.113▼	0.605▲	0.840	0.188▼	0.863°
Lab	0.034	0.087▲	0.127▲	0.213	0.528▲	0.582▲
Future	0.071	0.084°	0.170▲	0.243	0.408▲	0.519▲

Table 10.1: Experimental results for shot and program retrieval in the audiovisual archive, showing MAP scores for three query sets using three video search engines. ▲, ▼, and °, respectively indicate that a score is significantly better than, worse than, or statistically indistinguishable from the score using the Catalog video search engine.



(a) Experiment 1: 3x3 Shot Retrieval

(b) Experiment 2: 3x3 Program Retrieval

Figure 10.2: Experimental results for shot and program retrieval in the audiovisual archive, across three query sets and three video search engines. Note that performance more than doubles when using the Future search engine for shot retrieval on the Future queries.

10.3 Results

We now move on to the results of our experiments. The evaluation scores are summarized in Table 10.1. Additionally, Figure 10.2 highlights the different patterns in retrieval performance between query sets.

10.3.1 Experiment 1: 3x3 Shot Retrieval

The results for Experiment 1, i.e., shot retrieval with three video search engines (Catalog, Content and Future) using three query sets (Archive, Lab, Future), are presented in Figure 10.2a and Table 10.1 (columns 2–4).

The three query sets exhibit different sensitivity to the video search engines. The Archive queries attain significantly better performance using the Catalog video search engine than the Content video search engine, while the opposite is the case for the Lab queries. For Future queries, the performance of both of these search engines is similar.

The Future video search engine, which optimally combines the Catalog and Content engines, achieves significant improvements over the Catalog engine for all query sets. This effect is most marked for the Future queries, where performance more than doubles. Turning to the Archive queries, the increase in retrieval performance using the Future video search engine is relatively low at 12%.

We attribute the relatively good performance of the Catalog search engine for Archive queries to the nature of both the terms contained within the queries, and the process used to create relevance judgments. Recall that Archive queries and judgments are created by directly taking search and purchase information from the archive logs. The Catalog search engine frequently returns the correct program first, because the Archive queries are formulated in terms of the available archive catalog entries, which contain technical metadata unsuited for content-based video retrieval. For shot retrieval, when the correct program is returned first, all shots within it are also returned first. Now, turning to the creation of relevance judgements from purchase data, we found in Chapter 3 that 33% of the purchases in the archive are for entire programs. When an entire program is purchased, all of the shots within the program are judged as relevant, and within-program ordering does not make a difference. Therefore, when the Catalog engine returns all the shots from the correct program at the top of the result list, the Catalog search engine attains a high score.

In answer to CRQ 1, *What is the performance of content-based video retrieval when answering today's queries in the archive, and queries as they might be formulated in the archive of the future?*, content-based video retrieval alone is not enough to satisfy the needs of today's archive users. However, if future users state their information needs in content-based video retrieval terms (as is the case for the Future queries) then both search engines perform equally well. We gain the most when combining content-based video retrieval with retrieval using the catalog entries — which brings us to CRQ 2, *What can content-based video retrieval add to search performance when combined with current archive search capabilities?* Today's Archive queries, though less sensitive to content-based methods than other query sets, gain a significant performance increase by embedding content-based video retrieval into today's practice. After combination, tomorrow's Future queries gain even more, with performance more than doubling.

	Video Search Engine		
Query set	*Catalog*	*Content*	*Future*
Simulated	0.763	0.250 ▼	0.780 ▲

Table 10.2: Performance in MAP for Experiment 3; program retrieval for 2,190 simulated queries using three video search engines. ▲, ▼, and °, respectively indicate that a score is significantly better than, worse than, or statistically indistinguishable from the score using the catalog-based video search engine.

10.3.2 Experiment 2: 3x3 Program Retrieval

The results of Experiment 2, i.e., program retrieval with three video search engines using three query sets, are given in Figure 10.2b and Table 10.1 (columns 5–7).

As was the case for shot retrieval, the Archive queries are less responsive to the Content video search engine than the other two query sets. The Archive queries gain a high absolute MAP score of 0.840 with the Catalog search engine; the Content video search engine has a lower score of 0.188, and no significant improvement is gained by combining retrieval data sources in the Future video search engine. This is not surprising: once again, the poor performance of the Content search engine for these queries is due to the nature of the queries and judgments taken from the archive logs.The Lab and Future queries, on the other hand, perform better using the Content than the Catalog video search engine; this is to be expected as the queries were not created with reference to the catalog entries from the archive.

Returning to CRQ 3, *Can content-based video retrieval help those users that wish to retrieve not just shots, but entire programs?*, we can say that content retrieval does help to retrieve programs for tomorrow's Future queries, where visual information needs in the archive are formulated as multimedia queries. Queries taken directly from the archive logs did not prove sensitive to content-based video retrieval for program search: this is an artefact of the methodology used to create the queries and associated relevance judgments.

10.3.3 Experiment 3: Program Retrieval with Simulated Queries

The results for Experiment 3, i.e., program retrieval for 2,190 simulated purchase-query pairs, are shown in Table 10.2. The results for the simulated queries are similar to those for program retrieval with Archive queries as described in Section 10.3.2; the Catalog video search engine attains a higher performance than the Content engine. There is a 2% increase in performance when using the Future video search engine for retrieval.

The relatively high MAP score for the Catalog video search engine is to be expected, as the simulated queries have been generated from the catalog descriptions in the Archive Footage collection. Like the Archive queries, the query terms are sometimes taken from technical metadata that is not possible to locate using the Content-based search engine, for instance, 13% of the query terms are for the recording numbers contained in the catalog entries (see Table 4.1 on page 54). Indeed, for 30% of the queries the Catalog video search engine did not return any relevant results. However, in other cases the Content video search engine is at least as effective as the Catalog search engine, and for 19% of the queries, the Content video search engine gained an MAP score of 1, in other words, for these queries, the Content engine gave the simulated purchase the highest rank.

10.3.4 Experiment 4: Prioritizing Content Search

The results for Experiment 4, i.e., shot retrieval with three different content-based video retrieval methods, are shown in Table 10.3 and Figure 10.3. Notably, for the Future queries, there is no significant difference between the overall retrieval performances of transcript-based search, feature-based search and detector-based search. For the Lab queries, however, feature-based search and detector-based search significantly outperform transcript-based search. This can be explained by the visual nature of the Lab queries. These observations inform our answer to CRQ 4, *Which content-based video retrieval methods should be given priority for integration into the archive?* We give our answer using results from the Future queries, which are derived from logged archive searching behavior. For these queries, there is no significant difference between the three content-based video retrieval methods. On the basis of these results, it is not possible to recommend a single content-based video retrieval method, as they all give similar performance gains. Therefore other factors will need to be taken into account, such scalability, technological maturity, user acceptance, and ease of integration into the archive work-flow.

10.4 Conclusions and Recommendations

In this chapter, we studied the extent to which content-based video retrieval can enhance today's and tomorrow's retrieval performance in the audiovisual archive. To this end, we proposed an evaluation methodology tailored to the specific needs and circumstances of the archive. This methodology included three query set definitions, three state-of-the-art content-based and archive-based video search engines, and two challenging retrieval tasks that are grounded in a real-world audiovisual archive. We found that the greatest improvements in retrieval performance may be gained

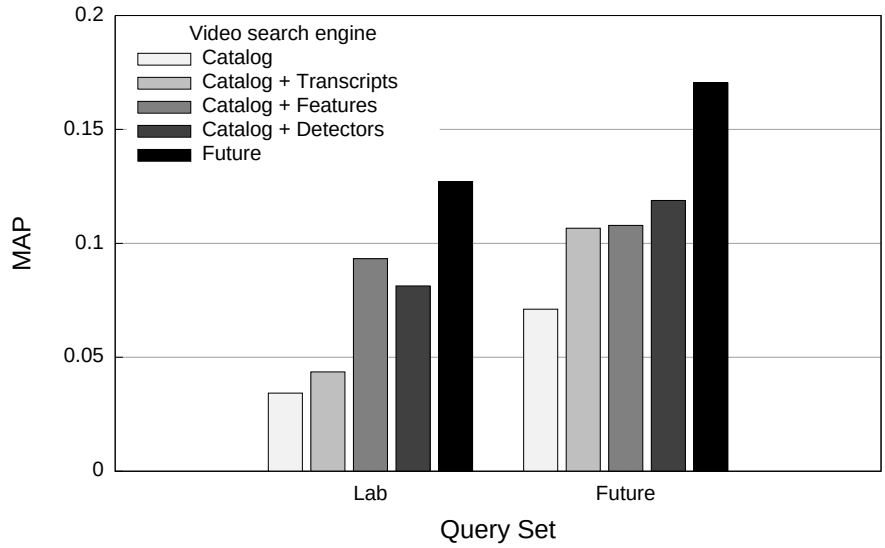


Figure 10.3: Shot retrieval MAP performance for the lab and future query sets, when combining retrieval results from the catalog-based video search engine with content-based video search methods.

	Content retrieval method		
Query set	*Transcript*	*Feature*	*Detector*
Lab	0.044 ▼▼	0.093 ▲°	0.081 ▲°
Future	0.107 °°	0.108 °°	0.119 °°

Table 10.3: Performance in MAP for Experiment 4; shot retrieval for two (multimedia) query sets using three different content-based video retrieval methods. ▲, ▼, and °, respectively indicate that a score is significantly better, worse, or statistically indistinguishable from the score of the remaining two content-based video retrieval methods, from left to right.

by combining content-based video retrieval with search using the manual catalog annotations that are maintained by the archive.

Our study was directed by four chapter-level research questions. In response to CRQ 1, *What is the performance of content-based video retrieval when answering today's queries in the archive, and queries as they might be formulated in the archive of the future?*, we found that for Future queries, content-based video retrieval outperformed traditional catalog-based video search engines of archives. To answer CRQ 2, *What can content-based video retrieval add to search performance when com-*

bined with current archive search capabilities?, we found that a catalog-based video search engine supplemented with content-based video retrieval potentially yields performance gains up to 270%. Our experiments with program-level retrieval indicate a positive answer to CRQ 3, *Which content-based video retrieval methods should be given priority for integration into the archive?*. We found that program retrieval with a content-based video search engine can potentially improve upon catalog-based search by up to 147%. Moreover, we evaluated program retrieval with a set of simulated purchase-query pairs, and found that content-based video retrieval alone was able to correctly identify the simulated purchase as the top result for 19% of the queries. When we examined individual content-based video retrieval methods in an attempt to answer CRQ 4, *Which content-based video retrieval methods should be given priority for integration into the archive?* we found that, based on retrieval experiments alone, none is to be preferred over the others as all three methods gave similar performance.

This brings us to our concluding recommendations. Our experiments have shown that content-based video retrieval aids the retrieval practice of the audiovisual archive. Hence, it is recommended that audiovisual archives invest in embedding content-based video retrieval into their work-flow. On the basis of video retrieval performance alone it is not possible to identify any single content-based video retrieval method to be given priority for inclusion into the archive, as they all perform equally well, therefore such prioritization should be based on other factors such as scalability, technological maturity, ease of integration into archive work-flow, and user acceptance. Yet the largest increase in retrieval performance is to be expected when transcript-based search is combined with a visual methodology using features and/or concept detectors. Audiovisual archives can not only profit from content-based video retrieval results, but also contribute to research by opening up their transaction logs and databases to study the valuable information inside. In this way content-based video retrieval and the audiovisual archive can mutually benefit from each other.

Chapter 11

Conclusion to Part II

In Part II of this thesis we have explored the use of automatic methods to improve search in audiovisual broadcast archives. To this end, we defined two evaluation collections as a basis for our experiments. The collections were obtained from existing benchmarks, and were therefore already associated with queries, judgments, videos, and automatically derived metadata. We additionally enriched one collection with a large-scale unified multimedia thesaurus of 450 visual concepts, which we created by linking existing concept lexicons to a single large-scale ontology. We enriched the second collection by adding query sets derived from the information needs recorded in the transaction logs of a real-world audiovisual archive.

Detector-based search relies on the automatic labelling of video fragments with respect to a lexicon of semantic concepts. However, these lexicons of concepts remain restricted to a few hundreds or thousands of entries, providing us with a limited vocabulary of labels with which to search. Thus, selecting those visual concept detectors that can best help us to answer a given information need is a challenging problem. We proposed two methods for creating concept selection benchmarks that map individual queries to appropriate entries in a concept lexicon. One benchmark creation method is based on human agreement achieved through focus group experiments, and the other is based on back-generation of concepts relevant to a query from a labelled training collection. We found that while the two benchmarks do not always agree, both methods can be used to successfully identify the best performing concept selection methods for retrieval.

With transcript-based search, we use the transcriptions of the spoken dialog of a video for retrieval. We investigated the temporal mismatch between the mention of an item in transcripts, and its appearance in video, and found that items are more likely to appear after they are mentioned in transcripts than they are to appear be-

fore they are mentioned. By modeling the characteristics of temporal redundancy and incorporating them in a retrieval framework, retrieval performance can be significantly improved.

What is the potential of content-based video retrieval to answer the information needs of professional users of an audiovisual archive? When the information needs recorded in the logs are formulated in terms of a content-based video retrieval system, such a system can outperform the catalog-based search on manual annotations traditionally used by audiovisual archives. Even larger performance gains can be obtained by combining the two sorts of systems. Based on our results, we believe that content-based video retrieval will be essential to the audiovisual archive of tomorrow.

Chapter 12

Conclusions to the Thesis

12.1 Answers to Research Questions

Part I of this thesis was devoted to deepening our understanding of the searcher in the audiovisual broadcast archive. We started our exploration by asking

> **RQ 1** *What kinds of content do media professionals search for, and in what manner do they search for it?*

Our transaction log analysis in Chapter 3 showed that, when considering the program titles and thesaurus entries of clicked results, professionals were searching for program names, person names, general subject words, locations, and other names, in that order. An additional analysis in Table 4.1 of Chapter 4 included extra data about matches between search terms and text in the technical metadata fields from purchased results, and showed that the professionals also search for document identifier codes and other technical metadata. In addition, date filters were also frequently used to select specific broadcast dates. From this we conclude that a significant portion of the searches that media professionals perform are for known items, where the searcher knows exactly which program they are looking for. Further, we examined the orders for audiovisual material that media professionals placed to the archive, and found that two thirds of them were not for entire programs, but rather for video stories and fragments.

We used the knowledge that we gained about the types of content that users were searching for in addressing our next research question,

> **RQ 2** *Can we recreate those searches by media professionals that result in*

purchases, and use them to create an artificial testbed for retrieval evaluation?'

Using the simulation framework that we developed in Chapter 4, we were able to create simulated pairs of queries and purchases that, when used for retrieval evaluation, approached the system rankings of real queries and purchases. One benefit of the simulator was that we were able to use it to create additional queries and relevance judgments for our archive-based test collection, where due to the limited number of programs only a small amount of real-world transaction log information was available.

In Part II of the thesis we studied how automatically generated content metadata could be used to improve the search for content in audiovisual broadcast archives. Upon a review of the literature, we asked questions about specific problems in content-based video retrieval, firstly,

> **RQ 3** *Given that search by visual concept detector is a valuable method for content-based video retrieval, how can we identify the correct visual concepts to use for a query?*

In answering this question we investigated the use of a concept selection benchmarks, which defined "correct" query-to-concept mappings on a per query basis, to assess whether a candidate set of selected concepts would do well for retrieval. We found that identifying the set of concepts to use for a query is no easy task; we compared two orthogonal, knowledge-intensive methods for concept selection benchmark creation, and found that neither of these approaches produced mappings that gave consistently the best performance when applied to the retrieval task. However, when applied to the task of assessing concept selection, we found that a benchmark created by mining collection knowledge produced assessments of concept selection that agreed with final retrieval performance. So to answer our research question, a set of benchmark query-to-concept mappings, created by mining an annotated collection, can be used to assess selection of the correct visual concepts for a query. This comes with the caveat that the concept of a "correct concept" is an elusive one when applied to this assessment task; we leave further investigation of this to future work.

Moving on to the problem of the temporal mismatch when searching with transcripts, we posed the question,

> **RQ 4** *Within a video broadcast, the same object may appear multiple times within and across the video and audio signal, for example being mentioned in speech and then appearing in the visual signal. How can this phenomenon be characterized, and can we model and use this characteristic so as to improve cross-stream retrieval of visual items using transcripts?*

We investigated the temporal mismatch, and the redundant occurrence of visual items in the video signal, in two domains: broadcast news, the audiovisual archive setting (with documentaries, current affairs magazine shows and educational programs). We characterized how occurrences of items were distributed across neighbouring shots, and used these distributions to inform a retrieval model that expands the transcripts associated with a shot with transcripts from neighboring shots. We found marked differences in the responsiveness of queries in the two domains to the document expansion process. The narrative structure of broadcast news, where news stories are presented in tight units one by one, meant that in broadcast news, queries benefited from expanding shots with transcripts from about ten adjacent shots, while in the archive domain retrieval performance kept improving as shots were expanded with transcripts from 60 adjacent shots. We feel this illustrates the difference in narrative structure across the two domains; broadcast news consists of clearly delineated short story units, while programs in the archive domain often do not contain clear story boundaries, and changes in visual content can be unpredictable.

Finally, we moved from our examination of problems in specific content-based video retrieval methods to study the effect of combining state-of-the-art content-based video retrieval methods with the retrieval capabilities of the archive, asking,

> **RQ 5** *What is the potential impact of content-based video retrieval in the audiovisual broadcast archive, taking into account both the needs of professional users, and the manually created data already present in the archive?*

Through our analysis of transaction logs from the Netherlands Institute of Sound and Vision we had already created two sets of queries that reflected the needs of media professionals. As the number of queries that was created through this process was relatively small, at a total of 65, we also included 2,190 simulated queries created by our best simulator. Our experiments showed that combining content-based video retrieval methods with search on manual catalog annotations from the archive resulted in significant increases in video retrieval performance. For content queries, performance more than doubled as compared to using the catalog annotations on their own.

12.2 Main Contributions

We group the main contributions of the thesis into two areas that correspond to the two parts of this thesis: understanding searchers in audiovisual broadcast archives, and improving search for content in audiovisual broadcast archives. The contributions are further structured per chapter.

Understanding Searchers

- Our first contribution in understanding searchers is a large-scale transaction log analysis of the electronic traces left behind by media professionals at a national audiovisual broadcast archive. As part of our contribution, we proposed a novel method for categorizing query terms, one which is applicable in environments where users search for fielded documents; query terms can be categorized by matching them to the field in which they appear in the clicked results. This allows a categorization of the long tail of query terms that are missed by typical frequency-based analyses.

- Our second contribution in this area is a framework for simulating logged user queries and purchases. Simulators developed within this framework can be used to create testbeds of artificial queries and purchases for evaluating different retrieval systems, and require only a collection of documents as input. The validation approach that we presented determines how well the rankings of retrieval systems on the simulator output correlate with the rankings of retrieval systems on a set of real logged searches and purchases. Although our experiments are performed in the setting of the audiovisual broadcast archive, the framework can be extended to other domains where users search for and purchase documents.

Improving Search for Content in the Archive

- Our first contribution for improving search for content in the archive is our study of two knowledge-intensive methods for determining which concepts from a limited vocabulary should be matched to a query for retrieval. The resulting query-to-concept mappings can be used to assess new sets of incoming concepts for the same query and predict their relative retrieval performance. The methods and insights that we have developed here can be used to help inform automatic concept selection algorithm automatic concept selection algorithmss for detector-based search. In addition, we feel they will also be of use for new problem areas where a limited number of concepts need to be matched to a query, for example in a setting where documents are tagged with terms from a small domain vocabulary.

- Our second contribution here is our investigation of the temporal mismatch and the redundancy of visual items in the video and audio signal (the latter manifested by proxy by their mention in transcript text). While this effect has been studied before, our analysis covered a greater temporal neighbourhood than other studies, and in addition was performed for hundreds of queries and concepts, in both the broadcast news and the audiovisual archive domains. Our

findings could help designers to improve the performance of transcript-based search systems for the two domains. In addition the characterizations of visual redundancy and the temporal mismatch could be insightful for researchers in the humanities who are interested in large-scale quantitative analyses of narrative structure.

- Our third contribution for this part is an analysis of ways in which automatically generated content metadata can be used to improve retrieval performance for the searches of media professionals. Our study coupled state-of-the-art content-based video retrieval methods to the practice of an audiovisual broadcast archive. The study can help inform policy makers at audiovisual archives who are facing the limitations of today's manual annotation practices and are considering incorporating content retrieval into their work-flow. In addition, both our findings and the resources that we have created can be used by researchers who wish to explore content-based video retrieval outside of the usual laboratory benchmark setting.

12.3 Future Directions

In this thesis we analyzed the transaction logs of an audiovisual archive to gain information about what media professionals were looking for in an audiovisual archive, and found that they were looking for (and buying) video fragments. We feel a valuable direction for future research lies in applying information from video purchases fragments to improve the search for video content. For example, in Chapter 7 we illustrated a search for "shots F16", that resulted in a purchase of several shots of material containing fighter jets from that broadcast. These keywords could be associated with the purchased shots, and returned in later searches for the same keyword. This task may not be as easy as it seems as at first glance; searchers in the audiovisual archive do not always provide clearly defined boundaries for their purchases (in other words, visual material relating to their keywords may be interspersed with non-related visual material), which brings us to another avenue for exploration.

We feel that the transaction log analysis presented in this thesis could be extended to study the potential of such logs for automatically creating large-scale sets of queries and relevance judgments for content-based video retrieval. The challenge here is automatically identifying the queries that are for visual content, and identifying which purchases and parts of purchases contain visually relevant material. This idea can be extended beyond the audiovisual broadcast archive; in online settings we are seeing the rise of *deep linking* for user-generated video. Video portals such as YouTube are enabling users to tag fragments within uploaded video. In addition, social commenting functions are enabled so that users can specify a time with

their comment. These types of annotations could also be used to create large-scale evaluation collections that approach the size of web data. Not only could such data be used for evaluation, it could also be used to learn query-specific concept detectors that are adapted to the environment in which the data is recorded.

Another area for future work lies in studying how automatically-generated metadata may be used to address queries that are not purely for visual information. In this thesis, and in content-based video retrieval studies in general, retrieval queries have so far been largely restricted to visual information needs. This is not the only dimension along which users may require audiovisual items. For example, users may require information along other information streams, such as quotations in the speech stream, or videos with specific emotional content. Or they may require iconic events, such as the marriage of a monarch. Identifying and addressing such information needs will increase the diversity of ways in which automatically generated content metadata can be applied to help searchers.

In this thesis we have seen that there is a demand in the audiovisual broadcast archive for access to video at the content level. In addition, we have demonstrated that retrieval performance for content queries can more than double when combining automatically generated content metadata with the manually created metadata already present in the archive. However, the test of acceptance will be for audiovisual broadcast archives to integrate automatically generated metadata into their video descriptions. Only then will we be able to determine the full potential for automatic methods to improve search in audiovisual broadcast archives.

Bibliography

[1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*, pages 19–26. ACM, 2006.

[2] R. Aly, D. Hiemstra, and A. de Vries. Reusing annotation labor for concept selection. In *Proceedings of the 8th International Conference on Image and Video Retrieval (CIVR '09)*, pages 1–8. ACM, 2009.

[3] A. Amir, J. Argillander, M. Berg, S. Chang, M. Franz, W. Hsu, G. Iyengar, J. Kender, L. Kennedy, C. Lin, et al. IBM research TRECVID-2004 video retrieval system. In *Proceedings of the 2nd TRECVID Workshop (TRECVID '04)*, 2004.

[4] K. Andreano, D. Streible, J. Horak, and P. Usai. The Missing Link: Content Indexing, User-Created Metadata, and Improving Scholarly Access to Moving Image Archives. *The Moving Image*, 7(2):82–99, 2007.

[5] L. H. Armitage and P. G. B. Enser. Analysis of user need in image archives. *Journal of Information Science*, 23(4):287–299, August 1997.

[6] J. A. Aslam and M. Montague. Models for metasearch. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01)*, pages 276–284, New York, NY, USA, 2001. ACM.

[7] L. Azzopardi. Query side evaluation: an empirical analysis of effectiveness and effort. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*, pages 556–563, New York, NY, USA, July 2009. ACM.

[8] L. Azzopardi, M. de Rijke, and K. Balog. Building simulated queries for known-item topics: an analysis using six European languages. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07)*, pages 455–462, New York, NY, USA, July 2007. ACM.

[9] M. Barbaro and T. Zeller. A face is exposed for AOL searcher no. 4417749. New York Times, August 9 2006.

[10] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[11] M. G. Brown, J. T. Foote, G. J. F. Jones, K. Sparck-Jones, and S. J. Young. Automatic content-based retrieval of broadcast news. In *Proceedings of the 3rd ACM International Conference on Multimedia (MULTIMEDIA '95)*, pages 35–43, San Francisco, USA, 1995. ACM Press.

[12] H. Brugman, V. Malaisé, and L. Gazendam. A web based general thesaurus browser to support indexing of television and radio programs. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC '06)*, pages 1488–1491, May 2006.

[13] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '00)*, pages 33–40, New York, NY, USA, 2000. ACM.

[14] M. Carman, M. Baillie, R. Gwadera, and F. Crestani. A statistical comparison of tag and query logs. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*, pages 123–130, Boston, USA, 2009. ACM.

[15] J. Carmichael, M. Larson, J. Marlow, E. Newman, P. Clough, J. Oomen, and S. Sav. Multimodal indexing of electronic audio-visual documents: a case study for cultural heritage data. In *Proceedings of the 6th International Workshop on Content-Based Multimedia Indexing (CBMI '08)*, pages 93–100, 2008.

[16] S. Carter, C. Monz, and S. Yahyaei. The QMUL system description for IWSLT 2008. In *Proceedings of the 5th International Workshop on Spoken Lanuage Translation (IWSLT '08)*, pages 104–107, 2008.

[17] D. Case. *Looking for information: A survey of research on information seeking, needs, and behavior*. Emerald Group Pub Ltd, 2007.

[18] S.-F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong. Videoq: an automated content based video search system using visual cues. In *Proceedings of the 5th ACM International Conference on Multimedia (MULTIMEDIA '97)*, pages 313–324, New York, NY, USA, 1997. ACM.

[19] S.-F. Chang, W. Hsu, W. Jiang, L. Kennedy, X. Dong, A. Yanagawa, and E. Zavesky. Columbia University TRECVID-2006 video search and high-level feature extraction. In *Proceedings of the 4th TRECVID Workshop (TRECVID '06)*, 2006.

[20] M. G. Christel. Establishing the utility of non-text search for news video retrieval with real world users. In *Proceedings of the 15th International Conference on Multimedia (MULTIMEDIA '07)*, pages 707–716, New York, NY, USA, September 2007. ACM.

[21] M. G. Christel and A. G. Hauptmann. The use and utility of high-level semantic features in video retrieval. In *Proceedings of the 4th International Conference on Image and Video Retrieval (CIVR '05)*, pages 134–144. Springer, 2005.

[22] T. Chua, S. Neo, K. Li, G. Wang, R. Shi, M. Zhao, H. Xu, Q. Tian, S. Gao, and T. Nwe. TRECVID 2004 search and feature extraction task by NUS PRIS. In *Proceedings of the 2nd TRECVID Workshop (TRECVID '04)*, 2004.

[23] T.-S. Chua, S.-Y. Neo, Y. Zheng, H.-K. Goh, Y. Xiao, S. Tang, and M. Zhao. TRECVID 2006 by NUS-I2R. In *Proceedings of the 4th TRECVID Workshop (TRECVID '06)*, 2006.

[24] B. Cole. Search engines tackle the desktop. *Computer*, 38(3):14–17, March 2005.

[25] E. Cooke, P. Ferguson, G. Gaughan, C. Gurrin, G. Jones, H. Le Borgne, H. Lee, S. Marlow, K. McDonald, M. McHugh, et al. TRECVID 2004 experiments in Dublin City University. In *Proceedings of the 2nd TRECVID Workshop (TRECVID '04)*, 2004.

[26] N. Craswell, D. Hawking, R. Wilkinson, and M. Wu. Overview of the TREC 2003 web track. In *Proceedings of the 12th Text Retrieval Conference (TREC '03)*, pages 78–92, 2003.

[27] N. Craswell, R. Jones, G. Dupret, and E. Viegas, editors. *Proceedings of the 2009 Workshop on Web Search Click Data (WSCD '09)*. ACM, New York, NY, USA, 2009.

[28] S. Cunningham and D. Nichols. How people find videos. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '08)*, pages 201–210. ACM New York, NY, USA, 2008.

[29] V. Dang and B. W. Croft. Query reformulation using anchor text. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM '10)*, pages 41–50. ACM, 2010.

[30] R. Datta, D. Joshi, J. Li, and J. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40(2):1–60, 2008.

[31] F. M. G. de Jong, T. Westerveld, and A. P. de Vries. Multimedia search without visual analysis: The value of linguistic and contextual information. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(3):365–371, 2007.

[32] O. de Rooij and M. Worring. Browsing video along multiple threads. *IEEE Transactions on Multimedia*, 12(2):121–130, 2010.

[33] O. de Rooij, C. G. M. Snoek, and M. Worring. Query on demand video browsing. In *Proceedings of the 15th International Conference on Multimedia (MULTIMEDIA '07)*, pages 811–814, New York, NY, USA, September 2007. ACM.

[34] O. de Rooij, C. G. M. Snoek, and M. Worring. Balancing thread based navigation for targeted video search. In *Proceedings of the 7th International Conference on Image and Video Retrieval (CIVR '08)*, pages 485–494, New York, NY, USA, 2008. ACM.

[35] J. Despres, P. Fousek, J.-L. Gauvain, S. Gay, Y. Josse, L. Lamel, and A. Messaoudi. Modeling northern and southern varieties of Dutch for STT. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH '09)*, Brighton, U.K., Sept. 2009.

[36] N. Dimitrova. Multimedia content analysis: The next wave. In *Proceedings of the 2nd International Conference on Image and Video Retrieval (CIVR '03)*, pages 415–420. Springer, 2003.

[37] C. Diou, G. Stephanopoulos, N. Dimitriou, P. Panagiotopoulos, C. Papachristou, A. Delopoulos, H. Rode, T. Tsikrika, A. De Vries, D. Schneider, et al. VITALAS at TRECVID-2009. In *Proceedings of the 7th TRECVID Workshop (TRECVID '09)*, 2009.

[38] R. Edmondson. *Audiovisual Archiving: Philosophy and Principles*. UNESCO, Paris, France, 2004.

[39] P. Enser. The evolution of visual information retrieval. *Journal of Information Science*, 34(4):531–546, 2008.

[40] P. Enser. Query analysis in a visual information retrieval context. *Journal of Document and Text Management*, 1(1):25–52, 1993.

[41] J. M. Ferryman, editor. *Proceedings of the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS '07)*. IEEE Computer Society, Rio de Janeiro, Brazil, 2007.

[42] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: The QBIC system. *Computer*, 28(9):23–32, 1995.

[43] J. Foote. An overview of audio information retrieval. *Multimedia Systems*, 7(1):2–10, 1999.

[44] E. A. Fox and J. A. Shaw. Combination of multiple searches. In *Proceedings of the 2nd Text REtrieval Conference (TREC-2)*, pages 243–252, Gaithersburg, USA, 1994.

[45] J. Garofolo, C. Auzanne, and E. Voorhees. The TREC spoken document retrieval track: A success story. In *Proceedings of the 9th Text Retrieval Conference (TREC '00)*, pages 107–130, 2000.

[46] M. Gordon. Evaluating the effectiveness of information retrieval systems using simulated queries. *Journal of the American Society for Information Science and Technology*, 41(5):313–323, 1990.

[47] J. Griffiths. *The computer simulation of information retrieval systems*. PhD thesis, School of Library, Archive and Information Studies, University College, London, UK, 1977.

[48] M. Halvey and J. Jose. The role of expertise in aiding video search. In *Proceedings of the 8th International Conference on Image and Video Retrieval (CIVR '09)*, pages 1–8. ACM, 2009.

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

[50] D. Harman. *The TREC test collection*, chapter 2, pages 21–52. TREC: Experiment and Evaluation in Information Retrieval, 2005.

[51] A. Haubold, A. Natsev, and M. Naphade. Semantic multimedia retrieval using lexical query expansion and model-based reranking. In *Proceedings of the International Conference on Multimedia and Expo (ICME '06)*, pages 1761–1764, 2006.

[52] C. Hauff, R. Aly, and D. Hiemstra. The effectiveness of concept based search for video retrieval. In *Proceedings of the Workshop Information Retrieval (FGIR '07)*, pages 205–212, 2007.

[53] A. Hauptmann, R. Yan, Y. Qi, R. Jin, M. Christel, M. Derthick, M.-Y. Chen, R. Baron, W.-H. Lin, and T. D. Ng. Video classification and retrieval with the informedia digital video library system. In *Proceedings of the 11th Text Retrieval Conference (TREC '02)*, 2002.

[54] A. Hauptmann, R. V. Baron, M. y. Chen, M. Christel, P. Duygulu, C. Huang, R. Jin, W. h. Lin, T. Ng, N. Moraveji, C. G. M. Snoek, G. Tzanetakis, J. Yang, R. Yan, and H. D. Wactlar. Informedia at TRECVID 2003: Analyzing and searching broadcast news video. In *Proceedings of the 1st TRECVID Workshop (TRECVID '03)*, 2003.

[55] A. G. Hauptmann. Lessons for the future from a decade of informedia video analysis research. In *Proceedings of the 4th International Conference on Image and Video Retrieval (CIVR '05)*, pages 1–10. Springer, 2005.

[56] A. G. Hauptmann, W. H. Lin, R. Yan, J. Yang, and M. Y. Chen. Extreme video retrieval: Joint maximization of human and computer performance. In *Proceedings of the 14th ACM International Conference on Multimedia (MULTIMEDIA '06)*, page 394. ACM, 2006.

[57] A. G. Hauptmann, R. Yan, W.-H. Lin, M. Christel, and H. D. Wactlar. Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news. *IEEE Transactions on Multimedia*, 9(5):958–966, 2007.

[58] J. He, C. Zhai, and X. Li. Evaluation of methods for relative comparison of retrieval systems based on clickthroughs. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09)*, pages 2029–2032, New York, NY, USA, 2009. ACM.

[59] M. Hertzum. Requests for information from a film archive: a case study of multimedia retrieval. *Journal of Documentation*, 59(2):168–186, March 2003.

[60] D. Hiemstra. *Using Language Models for Information Retrieval*. PhD thesis, University of Twente, Enschede, The Netherlands, 2001.

[61] K. Hofmann, M. de Rijke, B. Huurnink, and E. J. Meij. A semantic perspective on query log analysis. In *Working Notes for the CLEF 2009 Workshop*, September 2009.

[62] K. Hofmann, B. Huurnink, M. Bron, and M. de Rijke. Comparing click-through data to purchase decisions for retrieval evaluation. In *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10)*, pages 761–762, Geneva, July 2010.

[63] F. Hopfgartner and J. Jose. Evaluating the implicit feedback models for adaptive video retrieval. In *Proceedings of the 9th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR '07)*, page 331. ACM, 2007.

[64] F. Hopfgartner, D. Vallet, M. Halvey, and J. Jose. Search trails using user feedback to improve video search. In *Proceedings of the 16th ACM International Conference on Multimedia (MULTIMEDIA '08)*, pages 339–348. ACM, 2008.

[65] F. Hopfgartner, T. Urruty, P. Lopez, R. Villa, and J. Jose. Simulated evaluation of faceted browsing based on feature selection. *Multimedia Tools and Applications*, 47(3):631–662, 2010.

[66] G. Hripcsak and A. Rothschild. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3):296, 2005.

[67] W. Hsu, L. Kennedy, and S. Chang. Video search reranking via information bottleneck principle. In *Proceedings of the 14th ACM International Conference on Multimedia (MULTIMEDIA '06)*, page 44. ACM, 2006.

[68] M. Huijbregts, R. Ordelman, and F. de Jong. Annotation of heterogeneous multimedia content using automatic speech recognition. In *Proceedings of the 2nd International Conference on Semantics and Digital Media Technologies (SAMT '07)*, pages 78–90, Berlin, 2007. Springer Verlag.

[69] B. Huurnink and M. de Rijke. Exploiting redundancy in cross-channel video retrieval. In *Proceedings of the 9th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR '07)*, pages 177–186. ACM Press, September 2007.

[70] B. Huurnink and M. de Rijke. Term selection and query operations for video retrieval.

In *Proceedings of the 29th European Conference on Information Retrieval (ECIR '07)*, pages 708–711. Springer, April 2007.

[71] B. Huurnink and M. de Rijke. The value of stories for speech-based video search. In *Proceedings of the 6th International Conference on Image and Video Retrieval (CIVR '07)*, pages 266–271. Springer, July 2007.

[72] B. Huurnink, K. Hofmann, and M. de Rijke. Assessing concept selection for video retrieval. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval (MIR 2008)*, pages 459–466. ACM, October 2008.

[73] B. Huurnink, K. Hofmann, M. de Rijke, and M. Bron. Simulating searches from transaction logs. In *SimInt 2010: SIGIR Workshop on the Automated Evaluation of Interactive Information Retrieval*, July 2010.

[74] B. Huurnink, K. Hofmann, M. de Rijke, and M. Bron. Validating query simulators: An experiment using commercial searches and purchases. In *CLEF 2010: Conference on Multilingual and Multimodal Information Access Evaluation*, Padova, September 2010. Springer.

[75] B. Huurnink, L. Hollink, W. van den Heuvel, and M. de Rijke. The search behavior of media professionals at an audiovisual archive: A transaction log analysis. *Journal of the American Society for Information Science and Technology*, 61(6):1180–1197, June 2010.

[76] B. Huurnink, C. G. M. Snoek, M. de Rijke, and A. W. M. Smeulders. Today's and tomorrow's retrieval practice in the audiovisual archive. In *Proceedings of the 9th International Conference on Image and Video Retrieval (CIVR '10)*. ACM, ACM, July 2010.

[77] T. Ianeva, L. Boldareva, T. Westerveld, R. Cornacchia, D. Hiemstra, A. De Vries, and S. Valencia. Probabilistic approaches to video retrieval. In *Proceedings of the 2nd TRECVID Workshop (TRECVID '04)*, 2004.

[78] B. J. Jansen. The methodology of search log analysis. In B. J. Jansen, A. Spink, and I. Taksa, editors, *Handbook of Research on Web Log Analysis*, pages 99–121. Information Science Reference, 2008.

[79] B. J. Jansen and U. Pooch. A review of web searching studies and a framework for future research. *Journal of the American Society for Information Science and Technology*, 52 (3):235–246, 2001.

[80] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management*, 36 (2):207–227, 2000.

[81] B. J. Jansen, A. Spink, and J. O. Pedersen. The effect of specialized multimedia collections on web searching. *Journal of Web Engineering*, 3(3-4):182–199, 2004.

[82] B. J. Jansen, A. Spink, C. Blakely, and S. Koshman. Defining a session on web search engines. *Journal of the American Society for Information Science and Technology*, 58(6): 862–871, 2007.

[83] B. J. Jansen, D. L. Booth, and A. Spink. Determining the informational, navigational, and transactional intent of web queries. *Information Processing and Management*, 44 (3):1251–1266, 2008.

[84] Y. Jiang, J. Yang, C. Ngo, and A. Hauptmann. Representations of keypoint-based se-

mantic concept detection: A comprehensive study. *IEEE Transactions on Multimedia*, 12(1):42–53, 2010.

[85] T. Joachims. Optimizing search engines using clickthrough data. In *KDD '02*, pages 133–142, New York, NY, USA, 2002. ACM.

[86] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*, pages 154–161, New York, NY, USA, 2005. ACM.

[87] C. Jordan, C. Watters, and Q. Gao. Using controlled query generation to evaluate blind relevance feedback algorithms. In *JCDL '06*, pages 286–295, New York, NY, USA, 2006. ACM.

[88] C. Jörgensen and P. Jörgensen. Image querying by image professionals. *Journal of the American Society for Information Science and Technology*, 56(12):1346–1359, 2005.

[89] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):319–336, 2009.

[90] M. Kellar, C. R. Watters, and M. A. Shepherd. A field study characterizing web-based information-seeking tasks. *Journal of the American Society for Information Science and Technology*, 58(7):999–1018, 2007.

[91] L. S. Kennedy and S.-F. Chang. A reranking approach for context-based concept fusion in video indexing and retrieval. In *CIVR '07*, pages 333–340, New York, NY, USA, 2007. ACM.

[92] L. S. Kennedy, A. P. Natsev, and S.-F. Chang. Automatic discovery of query-class-dependent models for multimodal search. In *Proceedings of the 13th ACM International Conference on Multimedia (MULTIMEDIA '05)*, pages 882–891, New York, NY, USA, 2005. ACM.

[93] H. Keskustalo, K. Järvelin, A. Pirkola, T. Sharma, and M. Lykke. Test collection-based ir evaluation needs extension toward sessions–a case of extremely short queries. *Inf. Retr. Technology*, pages 63–74, 2009.

[94] W. Kraaij, A. Smeaton, P. Over, and J. Arlandis. TRECVID 2005 - an overview. In *Proceedings of the 3rd TRECVID Workshop (TRECVID '05)*, 2005.

[95] K. Krippendorff. *Content analysis: An introduction to its methodology*. Sage, Beverly Hills, CA, 1980.

[96] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01)*, pages 111–119, New York, NY, USA, 2001. ACM Press.

[97] G. Lakoff. *Women, Fire, and Dangerous Things*. University of Chicago Press, 1990.

[98] M. Larson, E. Newman, and G. Jones. Overview of VideoCLEF 2008: Automatic generation of topic-based feeds for dual language audio-visual contentideoclef 2008: Automatic generation of topic-based feeds for dual language audio-visual content. In *Working Notes for the CLEF 2008 Workshop*, Aarhus, September 2008.

[99] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2(1):1–19, 2006.

[100] X. Li, D. Wang, J. Li, and B. Zhang. Video search in concept subspace: a text-like paradigm. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval (CIVR '07)*, page 610. ACM, 2007.

[101] W.-H. Lin and A. G. Hauptmann. Which thousand words are worth a picture? Experiments on video retrieval using a thousand concepts. In *Proceedings of the International Conference on Multimedia and Expo (ICME '06)*, pages 41–44. IEEE, 2006.

[102] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[103] G. Marchionini. *Information seeking in electronic environments*. Cambridge Univ Pr, 1997.

[104] K. McDonald and A. F. Smeaton. A comparison of score, rank and probability-based fusion methods for video shot retrieval. In *Proceedings of the 4th International Conference on Image and Video Retrieval (CIVR '05)*, pages 61–70. Springer, 2005.

[105] T. Mei, Z.-J. Zha, Y. Liu, M. W. G.-J. Qi, X. Tian, J. Wang, L. Yang, and X.-S. Hua. MSRA att TRECVID 2008: High-level feature extraction and automatic search. In *Proceedings of the 6th TRECVID Workshop (TRECVID '08)*, 2008.

[106] E. Meij, M. Bron, L. Hollink, B. Huurnink, and M. de Rijke. Learning semantic query suggestions. In *Proceedings of the 8th International Semantic Web Conference (ISWC '09)*, pages 424–441, October 2009.

[107] G. A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38: 39–41, 1995.

[108] G. Mishne and M. de Rijke. A study of blog search. In *Proceedings 28th European Conference on Information Retrieval (ECIR '06)*, pages 289–301. Springer, April 2006.

[109] C. Monz and M. de Rijke. Shallow morphological analysis in monolingual information retrieval for Dutch, German, and Italian. In *Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems (CLEF '01)*, pages 262–277, London, UK, 2002. Springer-Verlag.

[110] M. Naphade, J. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE Multimedia*, 13(3):86–91, 2006.

[111] A. P. Natsev, A. Haubold, J. Tešić, L. Xie, and R. Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *Proceedings of the 15th International Conference on Multimedia (MULTIMEDIA '07)*, pages 991–1000, New York, NY, USA, 2007. ACM.

[112] R. B. Nelson. Kendall tau metric. In M. Hazewinkel, editor, *Encyclopaedia of Mathematics*, volume 3, pages 226–227. Springer, 2001.

[113] S.-Y. Neo, J. Zhao, M.-Y. Kan, and T.-S. Chua. Video retrieval using high level features: Exploiting query matching and confidence-based weighting. In *Proceedings of the 5th International Conference on Image and Video Retrieval (CIVR '06)*, pages 143–152, 2006.

[114] A. T. Nghiem, F. Bremond, M. Thonnat, and V. Valentin. ETISEO, performance evaluation for video surveillance systems. In *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 476–481, London, UK, 2007.

[115] C.-W. Ngo, Y.-G. Jiang, X. Wei, W. Zhao, F. Wang, X. Wu, and H.-K. Tan. Beyond semantic search: What you observe may not be what you think. In *Proceedings of the 6th TRECVID Workshop (TRECVID '08)*, 2008.

[116] P. Over, T. Ianeva, W. Kraaij, and A. Smeaton. TRECVID 2005 - an overview. In *Proceedings of the 3rd TRECVID Workshop (TRECVID '05)*. NIST, USA, 2005.

[117] P. Over, G. Awad, W. Kraaij, and A. F. Smeaton. TRECVID 2007 - an overview. In *Proceedings of the 5th TRECVID Workshop (TRECVID '07)*. NIST, USA, 2007.

[118] S. Ozmutlu, A. Spink, and H. C. Ozmutlu. Multimedia web searching trends: 1997-2001. *Information Processing and Management*, 39(4):611–621, 2003.

[119] E. Panofsky. *Studies in Iconology*. Harper and Row, New York, 1962.

[120] T. A. Peters. The history and development of transaction log analysis. *Library Hi Tech*, 11(2):41–66, 1993.

[121] C. Petersohn. Fraunhofer HHI at TRECVID 2004: Shot boundary detection system. In *Proceedings of the 2nd TRECVID Workshop (TRECVID '04)*, 2004.

[122] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, pages 275–281, New York, NY, USA, 1998. ACM Press.

[123] M. F. Porter. Dutch stemming algorithm. Retrieved January 25, 2010 from `http://snowball.tartarus.org/algorithms/dutch/stemmer.html`, 2010.

[124] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM '08)*, pages 43–52, 2008.

[125] S. Renals, T. Hain, and H. Bourlard. Interpretation of multiparty meetings: The AMI and AMIDA projects. In *Proceedings of the Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA '08)*, pages 115–118, Trento, Italy, 2008.

[126] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI '95)*, 1995.

[127] R. E. Rice and C. L. Borgman. The use of computer-monitored data in information science and communication research. *Journal of the American Society for Information Science and Technology*, 34(4):247–256, 1983.

[128] S. Robertson. On the history of evaluation in IR. *Journal of Information Science*, 34(4): 439–456, 2008.

[129] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, and M. Lau. Okapi at TREC. In *TREC*, pages 21–30, 1992.

[130] D. E. Rose and D. Levinson. Understanding user goals in web search. In *Proceedings of the 13th International Conference on World Wide Web (WWW '04)*, pages 13–19, New York, NY, USA, 2004. ACM Press.

[131] S. Rudinac, M. Larson, and A. Hanjalic. Exploiting result consistency to select query ex-

pansions for spoken content retrieval. In *Proceedings of the 32nd European Conference on Information Retrieval (ECIR '10)*, 2010.

[132] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[133] G. Salton, J. Allan, and C. Buckley. Approaches to passage retrieval in full text information systems. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '93)*, pages 49–58, New York, NY, USA, 1993. ACM.

[134] M. Sanderson. Test collection based evaluation of information retrieval systems. *Information Retrieval*, 4(4):247–375, 2010.

[135] C. J. Sandom and P. G. B. Enser. Virami - visual information retrieval for archival moving imagery. In *Proceedings of the 6th International Cultural Heritage Meeting (ICHIM '01)*, pages 141–152, Milan, Italy, 2001. Archives & Museum Informatics.

[136] R. Sargent. Verification and validation of simulation models. In *Proceedings of the 37th Winter Simulation Conference (WSC '05)*, page 143, 2005.

[137] A. Schutz and T. Luckmann. *The structures of the life-world*. Heinemann, London, 1974.

[138] R. Shaw. Royal society scientific information conference. *The American Statistician*, 2 (4):14–23, 1948.

[139] J. Simpson. March 2010 revisions - quarterly updates - Oxford English Dictionary, 2010. `http://www.oed.com/news/updates/revisions1003.html`. Retrieved 12 August 2010.

[140] A. Singhal and F. Pereira. Document expansion for speech retrieval. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*, pages 34–41, New York, NY, USA, 1999. ACM Press.

[141] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV '03)*, volume 2, pages 1470–1477, October 2003.

[142] A. Smeaton, P. Wilkins, M. Worring, O. de Rooij, T. Chua, and H. Luan. Content-based video retrieval: Three example systems from TRECVid. *International Journal of Imaging Systems and Technology*, 18(2-3):195–201, 2008.

[143] A. F. Smeaton. Techniques used and open challenges to the analysis, indexing and retrieval of digital video. *Information Systems*, 32(4):545–559, 2007.

[144] A. F. Smeaton and P. Over. TRECVID 2008: Search task (slides). In *Proceedings of the 6th TRECVID Workshop (TRECVID '08)*. NIST, USA, 2008.

[145] A. F. Smeaton, P. Over, and W. Kraaij. Trecvid: evaluating the effectiveness of information retrieval tasks on digital video. In *MULTIMEDIA '04*, pages 652–655, New York, NY, USA, 2004. ACM.

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

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

[148] J. R. Smith and S.-F. Chang. Visually searching the web for content. *IEEE Multimedia*, 4(3):12–20, 1997.

[149] J. R. Smith, S. Srinivasan, A. Amir, S. Basu, G. Iyengar, C.-Y. Lin, M. R. Naphade, D. B. Ponceleon, and B. Tseng. Integrating features, models, and semantics for TREC video retrieval. In *Proceedings of the 10th Text Retrieval Conference (TREC '01)*, 2001.

[150] S. W. Smoliar and H. Zhang. Content-based video indexing and retrieval. *IEEE MultiMedia*, 1(2):62–72, 1994.

[151] C. G. M. Snoek and A. W. M. Smeulders. Visual-concept search solved? *IEEE Computer*, 43(6):76–78, June 2010.

[152] C. G. M. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, 2005.

[153] C. G. M. Snoek and M. Worring. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 4(2):215–322, 2009.

[154] C. G. M. Snoek, J. C. van Gemert, T. Gevers, B. Huurnink, D. C. Koelma, M. van Liempt, O. de Rooij, K. E. A. van de Sande, F. J. Seinstra, A. W. M. Smeulders, A. H. C. Thean, C. J. Veenman, and M. Worring. The MediaMill TRECVID 2006 semantic video search engine. In *Proceedings of the 4th TRECVID Workshop (TRECVID '06)*, Gaithersburg, USA, November 2006.

[155] C. G. M. Snoek, M. Worring, J. M. Geusebroek, D. C. Koelma, F. J. Seinstra, and A. W. M. Smeulders. The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1678–1689, 2006.

[156] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the 14th ACM International Conference on Multimedia (MULTIMEDIA '06)*, pages 421–430, Santa Barbara, USA, October 2006.

[157] C. G. M. Snoek, I. Everts, J. C. van Gemert, J.-M. Geusebroek, B. Huurnink, D. C. Koelma, M. van Liempt, O. de Rooij, K. E. A. van de Sande, A. W. M. Smeulders, J. R. R. Uijlings, and M. Worring. The MediaMill TRECVID 2007 semantic video search engine. In *Proceedings of the 5th TRECVID Workshop (TRECVID '07)*, Gaithersburg, USA, November 2007.

[158] C. G. M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. T. Schreiber, and M. Worring. Adding semantics to detectors for video retrieval. *IEEE Transactions on Multimedia*, 9 (5):975–986, August 2007.

[159] C. G. M. Snoek, K. E. A. van de Sande, O. de Rooij, B. Huurnink, J. C. van Gemert, J. R. R. Uijlings, J. He, X. Li, I. Everts, V. Nedović, M. van Liempt, R. van Balen, F. Yan, M. A. Tahir, K. Mikolajczyk, J. Kittler, M. de Rijke, J.-M. Geusebroek, T. Gevers, M. Worring, A. W. M. Smeulders, and D. C. Koelma. The MediaMill TRECVID 2008 semantic video search engine. In *Proceedings of the 6th TRECVID Workshop (TRECVID '08)*, Gaithersburg, USA, November 2008.

[160] C. G. M. Snoek, M. Worring, O. d. Rooij, K. E. A. van de Sande, R. Yan, and A. G. Hauptmann. Videolympics: Real-time evaluation of multimedia retrieval systems. *IEEE MultiMedia*, 15(1):86–91, 2008.

[161] C. G. M. Snoek, K. E. A. van de Sande, O. de Rooij, B. Huurnink, J. R. R. Uijlings, M. van Liempt, M. Bugalho, I. Trancoso, F. Yan, M. A. Tahir, K. Mikolajczyk, J. Kittler, M. de Rijke, J.-M. Geusebroek, T. Gevers, M. Worring, D. C. Koelma, and A. W. M. Smeulders. The MediaMill TRECVID 2009 semantic video search engine. In *Proceedings of the 7th TRECVID Workshop (TRECVID '09)*, Gaithersburg, USA, November 2009.

[162] J. Tague and M. Nelson. Simulation of user judgments in bibliographic retrieval systems. *SIGIR Forum*, 16(1):66–71, 1981.

[163] J. Tague, M. Nelson, and H. Wu. Problems in the simulation of bibliographic retrieval systems. In *Proceedings of the 3d Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '80)*, pages 236–255, Kent, UK, 1980. Butterworth & Co.

[164] T. Tao, X. Wang, Q. Mei, and C. Zhai. Language model information retrieval with document expansion. In R. C. Moore, J. A. Bilmes, J. Chu-Carroll, and M. Sanderson, editors, *HLT-NAACL*. The Association for Computational Linguistics, 2006.

[165] O. Terris. There was this film about... the case for the shotlist. *Journal of Film Preservation*, 56:54—57, June 1998.

[166] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X.-S. Hua. Bayesian video search reranking. In *Proceedings of the 16th ACM International Conference on Multimedia (MULTIMEDIA '08)*, pages 131–140, New York, NY, USA, 2008. ACM.

[167] D. Tjondronegoro, A. Spink, and B. J. Jansen. A study and comparison of multimedia web searching: 1997-2006. *Journal of the American Society for Information Science and Technology*, 2009.

[168] T. Urruty, F. Hopfgartner, D. Hannah, D. Elliott, and J. Jose. Supporting aspect-based video browsing: analysis of a user study. In *Proceedings of the 8th International Conference on Image and Video Retrieval (CIVR '09)*, pages 1–8. ACM, 2009.

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

[170] W. van den Heuvel. Expert search for radio and television: a case study amongst Dutch broadcast professionals. In *Proceedings of the EuroITV2010 Workshop*, pages 47–50, 2010.

[171] W. van den Heuvel. Gebruikersonderzoek iMMix. Retrieved October 15, 2010, from `http://instituut.beeldengeluid.nl/index.aspx?ChapterID=8823`, Nederlands Instituut voor Beeld en Geluid, 2010.

[172] J. van Gemert, C. Veenman, and J. Geusebroek. Episode-constrained cross-validation in video concept retrieval. *IEEE Transactions on Multimedia*, 11(4):780–786, 2009.

[173] T. Volkmer, J. R. Smith, and A. P. Natsev. A web-based system for collaborative annotation of large image and video collections: an evaluation and user study. In *Proceedings of the 13th ACM International Conference on Multimedia (MULTIMEDIA '05)*, pages 892–901. ACM, 2005.

[174] E. Voorhees. The TREC question answering track. *Natural Language Engineering*, 7 (04):361–378, December 2001.

[175] E. Voorhees and D. Harman, editors. *TREC: Experiment and Evaluation in Information*

Retrieval. MIT Press, 2005.

[176] E. M. Voorhees. The philosophy of information retrieval evaluation. In *CLEF '01*, pages 355–370, London, UK, 2002. Springer-Verlag.

[177] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, pages 315–323, New York, NY, USA, 1998. ACM.

[178] H. D. Wactlar, M. G. Christel, Y. Gong, and A. G. Hauptmann. Lessons learned from building a terabyte digital video library. *IEEE Computer*, 32(2):66–73, 1999.

[179] D. Wang, X. Li, J. Li, and B. Zhang. The importance of query-concept-mapping for automatic video retrieval. In *Proceedings of the 15th International Conference on Multimedia (MULTIMEDIA '07)*, pages 285–288, 2007.

[180] Y. Wang, Z. Liu, and J. Huang. Multimedia content analysis using both audio and visual clues. *IEEE Signal Processing Magazine*, 17(6):12–36, November 2000.

[181] X.-Y. Wei and C.-W. Ngo. Ontology-enriched semantic space for video search. In *Proceedings of the 15th International Conference on Multimedia (MULTIMEDIA '07)*, pages 981–990. ACM, 2007.

[182] T. Westerveld. *Using generative probabilistic models for multimedia retrieval*. PhD thesis, University of Twente, Enschede, The Netherlands, 2004.

[183] P. Wilkins. *An investigation into weighted data fusion for content-based multimedia information retrieval*. PhD thesis, Dublin City University, Dublin, Ireland, 2009.

[184] P. Wilkins, R. Troncy, M. Halvey, D. Byrne, A. Amin, P. Punitha, A. Smeaton, and R. Villa. User variance and its impact on video retrieval benchmarking. In *Proceedings of the 8th International Conference on Image and Video Retrieval (CIVR '09)*, pages 1–8. ACM, 2009.

[185] P. Wilkins, A. F. Smeaton, and P. Ferguson. Properties of optimally weighted data fusion in cbmir. In *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10)*, pages 643–650, New York, NY, USA, 2010. ACM.

[186] R. Wilkinson. Effective retrieval of structured documents. In *SIGIR '94*, pages 311–317, NY, USA, 1994. Springer-Verlag New York, Inc.

[187] R. Wright. Annual report on preservation issues for European audiovisual collections. Deliverable PS_WP22_BBC_D22.4_Preservation Status_2007, BBC, 2007.

[188] G.-R. Xue, H.-J. Zeng, Z. Chen, Y. Yu, W.-Y. Ma, W. Xi, and W. Fan. Optimizing web search using web click-through data. In *Proceedings of the 13th International Conference on Information and Knowledge Management (CIKM '04)*, pages 118–126, New York, NY, USA, 2004. ACM.

[189] R. Yan. *Probabilistic Models for Combining Diverse Knowledge Sources in Multimedia Retrieval*. PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 2006.

[190] R. Yan and A. Hauptmann. Probabilistic latent query analysis for combining multiple retrieval sources. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*, page 331.

ACM, 2006.

[191] R. Yan and A. G. Hauptmann. A review of text and image retrieval approaches for broadcast news video. *Information Retrieval*, 10(4-5):445–484, 2007. ISSN 1386-4564.

[192] R. Yan, A. Hauptmann, and R. Jin. Negative pseudo-relevance feedback in content-based video retrieval. In *Proceedings of the 11th ACM International Conference on Multimedia (MULTIMEDIA '03)*, page 346. ACM, 2003.

[193] R. Yan, J. Yang, and A. G. Hauptmann. Learning query-class dependent weights in automatic video retrieval. In *Proceedings of the 12th ACM International Conference on Multimedia (MULTIMEDIA '04)*, pages 548–555, New York, NY, USA, 2004. ACM.

[194] J. Yang and A. G. Hauptmann. Exploring temporal consistency for video analysis and retrieval. In *Proceedings of the 8th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR '06)*, pages 33–42, New York, NY, USA, 2006. ACM Press.

[195] J. Yang, M. Y. Chen, and A. G. Hauptmann. Finding person X: Correlating names with visual appearances. In *Proceedings of the 3rd International Conference on Image and Video Retrieval (CIVR '04)*, pages 270–278. Springer, 2004.

[196] Y.-H. Yang and W. H. Hsu. Video search reranking via online ordinal reranking. In *Proceedings of the International Conference on Multimedia and Expo (ICME '08)*, pages 285–288. IEEE, 2008.

[197] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2):179–214, 2004.

[198] H. Zhang, J. Wu, D. Zhong, and S. Smoliar. An integrated system for content-based video retrieval and browsing. *Pattern Recognition*, 30(4):643–658, 1997.

Appendix A

Identifying Title and Thesaurus Terms in Queries

In Section 3.3.3, terms are identified by matching phrases and words within queries to metadata in clicked results. We include titles and thesaurus terms in the matching process, but exclude free text descriptions in the clicked results as these cannot be linked to any knowledge source. The resulting matched titles and entries are referred to as *title terms* and *thesaurus terms* respectively. The algorithm used to accomplish this is partially based on work by Snoek et al. [158] and outlined in Figure A.1. To summarize, a term is identified if: (1) at least one result is clicked during the search session in which the query is contained, (2) a candidate query phrase or word is contained in a program title or a thesaurus entry from a clicked result, and (3) there are not multiple candidate titles and/or entries conforming to the previous two conditions. As a result, not all terms can be identified, but coverage is extensive: of the $274,754$ non-empty queries $203,685$ were associated with at least one result click, and in turn $132,433$ queries were associated with at least one title term or thesaurus term.

Manual Identification of Query Terms To evaluate the quality and coverage of the automatically matched terms, an evaluation set was created. Here, three annotators manually identified the title terms and thesaurus terms contained in the queries from a randomly selected sample of 200 search sessions. The annotation procedure was set up as follows: the annotator was presented with the queries and clicked results from a search session, as well as a copy of the audiovisual thesaurus. The annotator was asked to identify the different title and thesaurus terms con-

Input
Queries (all keyword searches)
Catalog entries with titles and thesaurus terms
Result click data mapping queries in a session to clicked catalog entries in the same session
Output
Set of query terms (titles and thesaurus entries from the clicked results that match or contain phrases within the queries)
Step 0: Preprocessing
- associate each thesaurus entry in the collection with any synonyms that may be contained in the thesaurus
- *for each* query, title, and thesaurus entry in the collection
  - strip punctuation, diacritics
  - remove frequently occurring stop words
  - stem words using the Porter stemming algorithm for Dutch (123)
  - split compound words using a compound splitter adapted from (109)

Step 1: Selection of Candidate Terms
- *for each* query
  - associate the query with clicked terms in the form of titles and thesaurus entries contained in the clicked session results
  - *for each* clicked term
    - count how many times the clicked term appears in the clicked results

Step 2: Processing of Query Phrases
- *for each* query
  - create a set of all possible phrases within the query that maintain the sequence ordering. (The longest phrase will be the entire query, the shortest phrases will be the individual words contained within the query)
  - order phrases by length, so that the phrase with the most words comes first
  - *for each* query phrase
    - initialize empty set of matched terms
    - *if* the query phrase is identical to (exactly matches) at least one clicked term
      - add all identical clicked terms to the set of matched terms
    - *else, if* the query phrase is contained in (phrase matches) at least one clicked term
      - add all container clicked terms to the set of matched terms
    - *if* the set of matched terms contains exactly one term
      - **add the matched term to set of query terms**
      - remove all query phrases overlapping the current phrase from processing
      - *go to* next query phrase
    - *if* the set of matched terms contains more than one term
      - select the matched terms that occur the most frequently in clicked results
      - *if* there is single matched term occurs most frequently in clicked results
        - **add the single most matched term to set of query terms**
        - remove all query phrases overlapping the current phrase from processing
        - *go to* next query phrase
      - *if* multiple terms occur most frequently in clicked results, the query term is ambiguous
        - remove all query phrases overlapping the current phrase from processing
        - *go to* next query phrase
    - *go to* next query phrase
  - *go to* next query

Figure A.1: Process to identify title terms and thesaurus terms contained in user queries. Phrases within a query are matched to candidate titles and thesaurus entries from clicked results. When a match is found, all words from the query phrase are removed from the term identification process.

Table A.1: Evaluation of automatic term matching using precision and recall, based on a sample of 356 queries from 200 sessions. # *correctly matched* indicates the number of automatically matched terms that were also manually identified, # *matched* indicates the total number of automatically matched terms, # *correct* indicates the total number of manually identified terms.

Term source	Facet	# correctly matched	# matched	# correct	Precision	Recall
Title	n/a	108	114	157	0.95	0.69
Thesaurus	Genre	1	13	1	0.08	1.00
	Location	34	35	40	0.97	0.85
	Name	31	33	59	0.94	0.53
	Person	22	23	56	0.96	0.39
	Program Maker	3	7	7	0.43	0.43
	Subject	42	42	86	1.00	0.49
All terms	*n/a*	*241*	*267*	*406*	*0.90*	*0.59*

tained in a query using the clicked results and the audiovisual thesaurus. When a term could not be identified in the thesaurus or the titles of the clicked results, the annotator was asked to mark this separately as an *unknown* term. Our sample of 200 search sessions contained a total of 356 queries; our annotators identified a total of 157 title terms, 249 thesaurus terms, and 109 unknown terms. The unknown terms were not contained in the audiovisual thesaurus or the titles, and included 21 dates and 38 codes that were entered by the user into the keyword search. The remaining unknown terms consisted of concepts not present in the thesaurus, such as *night recording*. We investigated the agreement between the three annotators on a hold-out set of 30 sessions using Krippendorf's alpha [95]; we found the average pair-wise agreement to be 0.81.

Evaluation The performance of our automatic term identification method on the manually annotated sample of sessions is shown in Table A.1. Performance is measured in terms of precision (the number of automatically identified terms that were labelled as correct, divided by the total number of automatically identified terms), and recall (the number of automatically identified terms that were labelled as correct, divided by the total number of terms that were labelled as correct). Overall, the automatically identified terms are accurate, with on average nine out of ten of the identified terms being correct. Some types of terms are easier to identify correctly than others; title terms and thesaurus terms from the *Subject*, *Location*, and *Person* facets all have a precision of over 0.95, while thesaurus terms from the *Genre* and *Program Maker* facets have a precision of less than 0.50. Recall is similarly variable, with over a relatively large proportion of the manually identified *Location* and *Genre* terms being returned by the automatic method. Less than one in two of the

manually identified terms are identified for the *Person*, *Program Maker*, and *Subject* facets.

Appendix B

Visual Concepts in the Multimedia Thesaurus

Names of the 450 concepts contained in the unified multimedia thesaurus that is part of the Broadcast News collection.

Actor
Address Or Speech
Administrative Assistant
Adobehouses
Adult
Agent
Agricultural People
Aircraft
Aircraft Cabin
Airplane
Airplane Crash
Airplane Flying
Airplane Landing
Airplane Takeoff
Airport
Airport Building
Airport Or Airfield
Airport Terminal
Alley
Amusement Park
Animal
Animal Pens And Cages
Antenna
Apartment Complex
Apartments
Aqueduct
Ariel Sharon
Armed Person
Artillery
Asian People
Athlete
Attached Body Parts
Avalanche
Baby
Backpack
Backpackers
Baker
Balloons
Bank
Bar Pub
Barge
Barn
Barracks
Baseball Game
Basketball Game
Bathroom
Bazaar
Beach
Beards
Beggar
Bicycle
Bicycles
Bill Clinton
Bird
Blank Frame
Boat
Body Parts
Bomber Bombing
Boy
Bride
Bridges
Briefcases
Building
Bus
Business People
Cables
Camera
Canal
Candle
Canoe
Capital
Car
Car Crash
Car Racing Game
Cart Path
Cartoon Animation
Castle
Cats
Caucasians
Cave Inside
Celebration Or Party
Celebrity Entertainment
Cell Phones
Chair
Charts
Cheering
Cheerleader
Child
Cigar Boats
Cityscape
Civilian Person
Classroom
Clearing
Clock Tower

Clocks
Cloud
Cloverleaf
Coal Powerplants
Cockpit
Colin Powell
College
Commentator Or Studio Expert
Commercial Advertisement
Communications Tower
Computer Or Television Screen
Computers
Conference Buildings
Conference Room
Congressman
Construction Site
Construction Vehicles
Construction Worker
Control Tower - Airport
Cordless
Corporate Leader
Court
Courthouse
Crowd
Cul-de-sac
Cutter
Cycling
Dancing
Dark-skinned People
Daytime Outdoor
Dead Bodies
Demonstration Or Protest
Desert
Dining Room
Dirt Gravel Road
Ditch
Dog
Donald Rumsfeld
Donkeys
Drawing
Drawing And Cartoon Animation
Dredge Powershovel Dragline
Dresses
Dresses Of Women
Driver
Duo News Anchor Persons
Earthquake
Election Campaign
Election Campaign Address
Election Campaign Convention
Election Campaign Debate
Election Campaign Greeting
Embassy
Emergency Medical Response People
Emergency Room
Emergency Vehicles
Emile Lahoud
Empire State
Entertainment
Exiting Car
Exploding Ordinance
Explosion Fire
Eyewitness
Face
Factory
Factory Worker
Farms
Female Anchor
Female News Subject
Female Person
Female Reporter
Fields
Fighter Combat
Finance Busines
Fire Station
Fire Weapon
Firefighter
First Lady
Fish
Flag
Flag USA
Flood
Flowers
Flying Objects
Food
Football Game
Forest
Foxhole
Free Standing Structures
Freighter
Frigate
Funeral
Furniture
Gas Station
George Bush jr
George Bush sr
Girl
Glacier
Glass
Glasses
Golf Caddy
Golf Course
Golf Game
Golf Player
Government Leader
Grandstands Bleachers
Graphical Map
Graphics
Grass
Grassland
Graveyard
Greeting
Groom
Ground Combat
Ground Crew
Ground Vehicles
Group
Guard
Guest
Gym
Gymnastics
Hand
Handguns
Handshaking
Hangar
Harbors
Hassan Nasrallah
Head And Shoulder
Head Of State
Helicopter Hovering
Helicopters
High Security Facility
Highway
Hill
Horse
Horse Racing Game
Hospital Exterior
Hospital Interior
Host
Hotel
House
House Of Worship
Houseboat
Hu Jintao
Hunter
Individual
Indoor
Indoor Sports Venue
Industrial Setting
Infants
Insurgents
Interview On Location
Interview Sequences
Islands
Iyad Allawi
Jacques Chirac
Jail
John Edwards
John Kerry
Judge
Kitchen
Laboratory
Lakes
Landlines
Landscape
Laundry Room
Lawn
Lawyer
Logos Full Screen
Machine Guns
Male Anchor
Male News Subject
Male Person
Male Reporter
Marsh
Medical Personnel
Meeting
Microphones
Military Base
Military Buildings
Military Ground Vehicle
Military Personnel
Moonlight
Mosques
Motorcycle
Mountain
Muddy Scenes

Mug
Muslims
Natural Disaster
Network Logo
News Anchor Person
News Studio
News Subject Monologue
Newspaper
Night Vision
Nighttime
Non-uniformed Fighters
Non-us National Flags
Nuclear Powerplants
Observation Tower
Oceans
Office
Office Building
Officers
Oil Drilling Site
Oil Field
Old People
Outdoor
Outer Space
Overlayed Text
Parade
Parking Lot
Pavilions
Peacekeepers
Pedestrian Zone
People Crying
People Marching
People Walking
Person
Photographers
Pickup Truck
Pipelines
Pipes
Police
Police Security Personnel
Police Station
Politics
Pope
Porches
Power Plant
Power Transmission Line Tower
Powerlines
Powerplants
Press Conference
Priest
Prisoner
Processing Plant
Protesters
Raft
Railroad
Railroads
Rainy
Referees
Refugees
Religious Figure
Reporters
Residential Buildings
Rifles
Riot
River
River Bank
Road
Road Block
Road Overpass
Rocket Launching
Rocky Ground
Room
Rowboat
Rpg
Ruins
Running
Runway
Rv Farm
Sailboat
Scene Text
School
Science Technology
Scientists
Security Checkpoint
Ship
Shipyards
Shooting
Shopping Mall
Sidewalks
Singing
Single Family Homes
Single Person
Single Person Female
Single Person Male
Sitting
Ski
Skiing
Sky
Smoke
Smoke Stack
Snow
Soccer Game
Soldiers
Speaker At Podium
Speaking To Camera
Split Screen
Sport Games
Stadium
Standing
Steeple
Still Image
Stock Market
Store
Street Battle
Streets
Striking People
Studio With Anchorperson
Suburban
Subway Station
Suits
Sunglasses
Sunny
Supermarket
Swimmer
Swimming
Swimming Pool
Table
Talking
Tanks
Telephones
Television Tower
Tennis Game
Tent
Text Labeling People
Text On Artificial Background
Throwing
Ties
Toll Booth
Tony Blair
Tornado
Tower
Tractor Combine
Traffic
Tree
Tropical Settings
Truck
Tugboat
Tunnel
Underwater
United Nations Building
Urban
Urban Park
USA Government Building
Valleys
Vegetation
Vehicle
Violence
Volcano
Walking Running
Warehouse
Water Tower
Waterfall
Waterscape
Waterways
Weapons
Weather News
White House
Windows
Windy
Yasser Arafat

Appendix C

Queries

Identifiers and text descriptions for the queries contained in the Broadcast News and Archive Footage collections. English-language versions of the text descriptions are used where available.

Broadcast News collection - all 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 W. 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
8001. Military formations engaged in tactical warfare, or part of a parade
8002. Government or Civilian leaders at various locations such as press conference, indoors, outdoors, meeting other leaders, addressing crowds, rallies, in parliament or legislative buildings, at photo opportunities etc.
8004. Crowds protesting on streets in urban or rural backgrounds with or without posters/banners

etc.
8006. Funeral Procession or Scenes from a funeral or from a cemetery/crematorium/burial site with participants chanting slogans, and/or armed militia and/or masked people present, people carrying pictures of the dead
8007. People on street expressing sorrow by crying, beating their chests, chanting
8008. Military vehicles or helicopters
8011. Police firing weapons
8012. People touching a coffin
8016. Armed guards at checkpoints with barricade on roads
8017. Injured or dead people lying on the ground in any location such as in front of a mosque, on a street, in open grounds, in water etc
8018. Presidential Candidates
8019. Vice-presidential Candidates
8020. Indoor Debate with Speakers at Podium
8021. Town-hall Style Gathering
8022. U.S. Maps depicting the electoral vote distribution (blue vs. red state)
8027. Indoor scene with speaker addressing audience waving flags and cheering
8029. Person greeting people or crowd
8030. Two people on stage in a debate
8031. People posing for pictures with cameras flashing
8034. Soldier sniping at target
8036. Armed men on the city streets
8039. Tanks rolling on streets
8040. Tanks rolling in desert
8041. Armed uniformed soldiers walking on city lanes
8047. Cars burning on city streets or in the desert. May also have overturned cars by the side of roads
8052. Person People not in uniform firing weapons
8053. Armed Soldiers firing weapons
8059. Person or people not in uniform, firing weapon and hiding behind wall of house or building
8070. Armored Vehicles driving through barren landscapes
8074. Refugee Camps with women and children visible
8080. Empty Streets with buildings in state of dilapidation
8087. Man firing shoulder fired missile in air
8091. Armed Guards standing outside large buildings
8093. Protests turning violent with people throwing missiles, burning objects and clashing with armed military personnel
8099. Military meeting in an indoor setting with flag visible
8100. Vehicles with flags passing on streets
8101. An open air rally with a high podium and people attending
8103. Rebels with guns on streets or in jeeps
8107. People on the streets being interviewed by a reporter speaking into a microphone
8109. Scenes of battle between rebels and military in urban setting
8119. Destroyed aircrafts and helicopters
8121. Demonstrators marching on streets with banners and signs against military brutality
8128. Rebels brandishing and firing weapons in the air

Archive Footage collection - Archive queries
104822. shots f16, saab vliegtuig shots
105957. td43855
123568. zembla 03-06-1999
134947. westerbork zw/w mxf
155376. 111920
168932. zoo haanstra
173107. fotograferen, kiekjes maken, kiekjes mxf mxf mxf
178368. kerstmis
185943. nieuws uit de natuur
186097. den helder, nff, gouden kalf, pipo nps nps nps nps nps nps nps
197929. oranjemars
204112. bohr, niels bohr, bohr oppenheimer, bohr heisenberg, bohr noorderlicht
2230. noorderlicht
233476. alle dieren tellen mee
234800. klokhuis wapenaar
24877. roma zigeuners
267950. thuis in het universum, het beeld van de dolfijn, het multi-universum, multi-universum
26969. school tv weekjournaal
275011. glas haanstra, haanstra
278588. maritiem, water nps nps
283047. papa godett andere tijden
283349. andere tijden amateurbeelden oorlog, andere tijden tweede wereldoorlog amateurbeelden
289333. 166250
30886. td43855
31094. zoo haanstra

41282. kamp westerbork, kamp vught, kamp sobibor
44680. 874
47903. td37144
53399. sjah
60256. zoo haanstra, haanstra dierentuin
64597. weiland koeien, weiland schaoen, weiland schapen, schapen wei mxf mxf mxf mxf
66109. doodknuppelen zeehond
78951. marco gerris nps
81811. love boat, missionarissen, missionaris, zendeling*, polygoon zendeling, polygoon afrika, polygoon missie, polygoon journaal missie mxf mxf mxf mxf mxf mxf
84322. mensen in een dierentuin, mensen in een dierentuin bert haanstra
94314. zembla verdacht van beursfraude

Archive Footage collection - Future queries
3001. Find shots of a set table, with cutlery and at least one plate visible.
3002. Find shots of mushrooms in the forest.
3003. Find shots of a welder at work.
3004. Find shots of a burning cigarette.
3005. Find shots of a vendor behind the counter of a store
3006. Find shots of a symphony orchestra or chamber orchestra.
3007. Find shots of an elderly woman, with gray or white hair and lots of wrinkles.
3008. Find shots of a field with cows.
3009. Find shots of the working area in a factory.
3010. Find shots of a slum.
3011. Find shots of an old man (gray haired or balding, with many wrinkles) with one or more children.
3012. Find shots of shopping carts.
3013. Find shots of ice skaters with (a part of) their legs visible.
3016. Find shots of religious objects such as gold crosses and statues of saints. Objects may come from all religions.
3017. Find shots with a close-up of an insect on a leaf.
3018. Find shots of a hut with a thatched roof.
3019. Find shots of a computer animation of a process.
3020. Find shots with a patient in an operating room.
3021. Find shots of a large pasture with a house or farm building in the background.
3022. Find shots with the sun or moon shining through the silhouette of a tree or branches.
3024. Find shots of a small child in a child's chair.
3025. Find shots of a construction site.
3026. Find shots of a naked torso.
3028. Find shots of Princess Maxima and/or Prince Willem Alexander.
3030. Find shots of a meeting in the Lower House of parliament.
3032. Find shots of Queen Beatrix.
3034. Find shots of a group of press photographers.
3035. Find shots of a blonde woman.
3036. Find shots of Job Cohen, the mayor of Amsterdam.

Archive Footage collection - Lab queries
0200. Find shots of hands at a keyboard typing or using a mouse.
0201. Find shots of a canal, river, or stream with some of both banks visible.
0203. Find shots of a street market scene.
0204. Find shots of a street protest or parade.
0205. Find shots of a train in motion.
0206. Find shots with hills or mountains visible.
0207. Find shots of waterfront with water and buildings.
0210. Find shots with one or more people walking with one or more dogs.
0211. Find shots with sheep or goats.
0217. Find shots of a road taken from a moving vehicle through the front windshield.
0219. Find shots that contain the Cook character in the Klokhuis series.
0220. Find grayscale shots of a street with one or more buildings and one or more people.
0221. Find shots of a person opening a door.
0222. Find shots of 3 or fewer people sitting at a table.
0223. Find shots of one or more people with one or more horses.
0224. Find shots of a road taken from a moving vehicle, looking to the side.
0225. Find shots of a bridge.
0226. Find shots of one or more people with mostly trees and plants in the background; no road or building visible.
0227. Find shots of a person's face filling more than half of the frame area.
0228. Find shots of one or more pieces of paper,

each with writing, typing, or printing it, filling more than half of the frame area.
0229. Find shots of one or more people where a body of water can be seen.
0230. Find shots of one or more vehicles passing the camera.
0231. Find shots of a map.
0232. Find shots of one or more people, each walking into a building.
0233. Find shots of one or more black and white photographs, filling more than half of the frame area.
0234. Find shots of a vehicle moving away from the camera.
0235. Find shots of a person on the street, talking to the camera.
0236. Find shots of waves breaking onto rocks.
0237. Find shots of a woman talking to the camera in an interview located indoors - no other people visible.
0238. Find shots of a person pushing a child in a stroller or baby carriage.
0239. Find shots of one or more people standing, walking, or playing with one or more children.
0240. Find shots of one or more people with one or more books.
0241. Find shots of food and/or drinks on a table.
0242. Find shots of one or more people, each in the process of sitting down in a chair.
0243. Find shots of one or more people, each looking into a microscope.
0244. Find shots of a vehicle approaching the camera.
0245. Find shots of a person watching a television screen - no keyboard visible.
0246. Find shots of one or more people in a kitchen.
0247. Find shots of one or more people with one or more animals.
0248. Find shots of a crowd of people, outdoors, filling more than half of the frame area.
0249. Find shots of a classroom scene.
0250. Find shots of an airplane exterior.
0251. Find shots of a person talking on a telephone.
0252. Find shots of one or more people, each riding a bicycle.
0253. Find shots of one or more people, each walking up one or more steps.
0254. Find shots of a person talking behind a microphone.
0255. Find shots of just one person getting out of or getting into a vehicle.
0256. Find shots of one or more people, singing and/or playing a musical instrument.
0257. Find shots of a plant that is the main object inside the frame area.
0258. Find shots of one or more people sitting outdoors.
0259. Find shots of a street scene at night.
0260. Find shots of one or more animals - no people visible.
0261. Find shots of one or more people at a table or desk, with a computer visible.
0262. Find shots of one or more people in white lab coats.
0263. Find shots of one or more ships or boats, in the water.
0264. Find shots of one or more colored photographs, filling more than half of the frame area.
0266
0265. Find shots of a man talking to the camera in an interview located indoors - no other people visible.
0266. Find shots of more than 3 people sitting at a table.
0267. Find shots with the camera zooming in on a person's face.
0268. Find shots of one or more signs with lettering.
9001. Find shots of a vehicle travelling past a road sign.
9002. Find shots of people eating.
9003. Find shots of a discussion with more than one people wearing glasses.
9004. Find shots of at least one boat or any watercraft (including submarines, canoes, etc.)
9005. Find shots of traffic signs.
9006. Find shots showing interior or exterior building doors with at least half the height and half the width subtended by glass.
9007. Find shots of numbers overlayed on the screen. The height of the numbers should be larger than at least half of the screens.
9008. Find shots of one or more people reading a book, a magazine or a newspaper.
9009. Find shots of a person playing a guitar.
9010. Find shots of people walking in a forest.
9011. Find shots of the interior or exterior of a church.
9012. Find shots of apes or monkeys.

Samenvatting

Documentairemakers, journalisten, nieuwsredacteuren en andere media professionals hebben regelmatig audiovisueel archiefmateriaal nodig voor nieuwe producties. Zo kan een nieuwsredacteur voor het journaal internationaal beeldmateriaal willen gebruiken, en kan de maker van een documentaire over tradities rond kerst op zoek zijn naar beelden van kerstbomen uit verschillende decennia. Belangrijke bronnen voor beeldmateriaal zijn audiovisuele archieven, die audiovisueel materiaal bewaren en beheren.

In dit proefschrift bestuderen wij het zoeken in audiovisuele archieven. Onze aanpak is tweeledig, en het proefschrift bestaat daarom ook uit twee delen. In Deel I wordt de zoeker bestudeerd, met onder andere een analyse van de handelingen van media professionals op de online zoekinterface van een nationaal audiovisueel archief. Hierbij kijken wij zowel naar hun zoekopdrachten als hun aankopen. Om het gedrag van media professionals te kunnen modeleren doen wij op basis van deze analyse een simulatie-experiment. Hierin onderzoeken wij verschillende methodes voor het modelleren van de zoekopdrachten en aankopen in het archief.

Met dit verbeterde inzicht in de zoeker bestuderen wij in Deel II het zoeken. Wij bestuderen onder andere hoe state-of-art methodes voor het automatisch genereren van metadata bij videomateriaal gebruikt kunnen worden om het zoeken naar audiovisuele fragmenten te verbeteren, specifiek in de context van audiovisuele archieven. Wij ontwikkelen ook testcollecties voor het evalueren van nieuwe zoekalgoritmes. Deze zijn deels gebaseerd op informatie over zoekers die in Deel I vergaard is. De testcollecties vormen de basis voor experimenten gericht op het oplossen van specifieke problemen die spelen wanneer er wordt gezocht met automatisch gegenereerde beschrijvingen van videomateriaal. Tot slot voegen wij zowel automatisch gegenereerde als handmatig gecreëerde beschrijvingen samen uit meerdere bronnen, en bestuderen hun potentieel om zoeken in audiovisuele archieven te verbeteren.

Summary

Documentary makers, journalists, news editors, and other media professionals routinely require previously recorded audiovisual material for new productions. For example, a news editor might wish to reuse footage shot by overseas services for the evening news, or a documentary maker describing the history of the Christmas tradition might desire shots of Christmas trees recorded over the decades. Important sources for reusable broadcasts are audiovisual broadcast archives, which preserve and manage audiovisual material.

In this thesis we study search in audiovisual broadcast archives. Our approach is twofold, and accordingly the thesis is structured in two parts. Part I is dedicated to studying the *searcher*, and includes an analysis of the actions of media professionals on the online search interface of a national broadcast archive. We characterize their actions both in terms of their searches and in terms of their purchases. In order to model the behavior of media professionals we follow this analysis with a simulation experiment. Here we investigate different methods for modelling the searches and purchases recorded in the archive.

Having gained a better understanding of the searcher, in Part II we move on to study *search*. In particular, we investigate how state-of-art methods for automatically generating content metadata from video may be used to improve the search for audiovisual fragments, in the context of the audiovisual broadcast archive. We use data from the searchers that we studied in Part I to help define new test collections for retrieval evaluation. These are used as the basis for experiments aimed at solving specific problems that are faced when searching with automatically generated descriptions of video content. Finally, we combine multiple sources of retrieval information, both automatically generated and manually created, and investigate their potential impact on search in the archive.