A Multidisciplinary Approach to Unlocking Television Broadcast Archives

Laura Hollink[1] Bouke Huurnink[2] Michiel van Liempt[2]
Johan Oomen[3] Annemieke de Jong[3] Maarten de Rijke[2]
Guus Schreiber[1] Arnold Smeulders[2]

1: VU University Amsterdam, de Boelelaan 1081a, 1081 AH Amsterdam

2: ISLA, University of Amsterdam, Science Park 107, 1098 XG Amsterdam

3: Netherlands Institute for Sound and Vision, Sumatralaan 45, 1200 BB Hilversum

Abstract

Audiovisual material is a vital component of the world's heritage but it remains difficult to access. With the Netherlands Institute for Sound and Vision as one of its partners, the MuNCH project aims to investigate new methods for improving access to a wide range of audiovisual documents. MuNCH brings together three research fields: multimedia analysis, language technology and semantic technologies.

Within the MuNCH project we have investigated several combinations of these fields. We have compared text matching, ontology querying, and semantic visual querying as methods to translate a multimedia query to the vocabulary of the retrieval system. In addition, we have investigated how users make such a translation, and have used this as a benchmark to create automatic methods. We have used multimedia technology to automatically detect objects and scenes as they occur in video, and made use of language technology to exploit automatic transcriptions of speech. We have enriched the Sound and Vision thesaurus that is used to annotate the TV programmes in order to provide a user with a wider range of search results.

In order to verify the results of the project against real user needs, MuNCH has participated in the creation of a logging system which monitors the usage of the Sound and Vision catalogue system. Insights in the needs of real users will be used as input for all three of MuNCH's research strands.

1 Introduction

Digital video exploded onto our screens at the beginning of the new millennium. Television broadcasts are now commonly transmitted digitally via ether, cable, satellite, or online. Furthermore, the traditional repositories of audiovisual material — film and television broadcast archives — are now rapidly digitising the analog multimedia that lies locked within their vaults, with a profound effect on the demand for access to archive content (Edmondson 2004). Digitisation is only the first step. Semi-automatic annotation will become a pre-condition for maintaining high service levels, as manually created descriptions of video content do not suffice when scaling to increasingly large collections.

Now that content has been digitised, the users of public television broadcast archives can be provided with online and on-demand access to video. They no longer need to physically come to the archive and consult with an archive customer service agent to search for, view, and purchase video material. This increases ease of access. However, the burden of search is now placed squarely on the shoulders of the users, who may lack the specialised knowledge of the archive contents that the archive's customer service agent has. With the burden of search transferred to the non-expert, efficient access becomes an important factor in achieving a satisfactory user experience.

In this paper we describe the Multimedia aNalysis for Cultural Heritage (MuNCH) project, an interdisciplinary research effort aimed at investigating new methods for providing access to the digital video collections contained within public broadcast archives. This research is placed within the context of the Netherlands Institute for Sound and Vision (also referred to as "Sound and Vision", or "the Institute"). As one of the biggest audio-visual archives in Europe, Sound and Vision fulfills a unique role in the preservation and disclosure of Dutch culture. It stores over 700,000 hours of audiovisual material, and receives more material every day. 8,000 hours of digitally born television programmes are ingested in the asset management system annually, and an additional 20,000 hours of legacy material is digitised every year. The creative industry (broadcasters, newsmakers etc.) of the Netherlands relies on the institute for reuse of audiovisual material. The general public as well as teachers and students are able to access this part of the nation's heritage online through tailored services. By hosting the MuNCH project, Sound and Vision prepares to meet the needs of their customers now and in the future of the digital era. As the cultural heritage partner in the project, Sound and Vision provides video data, manual annotations created by their cataloguers, and retrieval use case scenarios. With that they set the stage for MuNCH's scientific work, one that is well grounded in current practice.

Providing continuous access to an archive as large as the Netherlands Institute for Sound and Vision is an ambitious goal. To this end the Institute employs a team of professional cataloguers, who create manual descriptions of the archive's broadcasts as they are acquired. These manual descriptions are used as the basis of the Institute's search engine. They can be rich and highly detailed, however extensive catalogue entries can only be created for a small fraction of the new videos that come in, especially with the video ingestion rate increasing rapidly. In addition, the cataloguers are instructed to create descriptions specifically for professional users, resulting in annotation bias.

Automation is needed. However, the fundamental obstacle in automatic annotation is the semantic gap (Smeulders et al. 2000) between the digital data and their semantic interpretation. To tackle the tough problem of semantic video retrieval, we put to use all available sources; we combine visual data with textual information such as speech recognition transcripts. Explicit knowledge about the domain of broadcast television is used to form a bridge between the digital data and the vocabulary of a user. To this end, MuNCH brings together three research fields: multimedia analysis, language technology and semantic technologies. The synergy between these fields can potentially lead to results that could not have been accomplished by each field separately.

In this paper we present the cooperative effort of the MuNCH disciplines in unlocking the information contained in the archives of the Netherlands Institute for Sound and Vision. We start by introducing the users of the archives in and then outline our proposed system for processing archival content. Next we present contributions of the multimedia analysis, language technology, and semantic technology disciplines, as well as cooperations between them. We conclude with a discussion of future directions for unlocking audiovisual broadcast archives.

2 Users of the Television Broadcast Archive

Every day thousands of people search the archives of the Netherlands Institute for Sound and Vision. Many of these users are television professionals, such as producers, broadcasters and directors, that are looking for reusable shots. A second user group is people from science and education: historians and communication scientists study the material in the archives; students and teachers from primary and secondary schools use the material to illustrate or clarify their lessons. A third user group consists of people with a general interest in a programme or a topic, who use the Institute's website to access the archives from their homes, their jobs, or as part of a visit to the media museum that is located within the building of Sound and Vision.

These diverse user groups have a broad range of search needs. Queries can be on the level of what the programme is about, what can be seen in the shots, or both; they can be targeted towards broad categories of topics or genres, a specific programme or a single shot. Some users know exactly what they are looking for, while others have only a vague idea. The needs of television professionals relate to the genre and developmental stage of the programmes they make. A journalist who searches for a shot to illustrate an item in tomorrow's news bulletin may only have time to quickly scan the descriptions of a few programmes for a usable shot, while a documentary maker may have time to view multiple complete programmes before selecting a shot with the right content, atmosphere and aesthetic qualities. Years of experience at the customer service department of Sound and Vision have led to the following broad categorisation of user queries:

Known item queries "The item about health care in the NOS news broadcast of the 15th of June 2008", "The documentary by Henk de By about the Dutch painter Melle"

Subject queries General areas of interest : “all programmes about the Dutch economy”, Recognised areas of interest : “housing problems of Spanish immigrants in Amsterdam during the sixties”

Sequences, shots and quotes Specific: “shots of George Bush announcing war with Iraq”, General: “shots of sunsets”; “shots of Newfoundland”.

Information about the clients and their needs is essential for setting the requirements of a video retrieval system. Sound and Vision can use this information to fine-tune their services to the needs of their customers. For scientific goals it can provide a source of real-life use cases to target our research. To increase knowledge of user needs, MuNCH has helped Sound and Vision design a query logging module that registers user behavior at the website of Sound and Vision. The website provides access to the archives by means of a search interface. Users can search by keyword or select a thesaurus term specifying broadcaster, subject, genre, name, etc. programmes in the result list can be viewed and ordered online. All actions of the users of this website are logged and stored in the query logging database. For example, a record is made of every time a user searches for a keyword, clicks on an item in the result list, refines a search using thesaurus terms, previews a video of the programme, bookmarks a programme, puts a programme in his or her shopping cart, etc. In total, 41 actions are distinguished. In addition, we can track users over time, to record which consecutive actions a user performs during a session. This information on user behavior provides a wealth of information about how users search and what they search for.

A first analysis of the user logs results in frequency information; it gives an answer to the question “How often is a particular (type of) query posed?” This is valuable information for both Sound and Vision and the video retrieval community, since it allows for the identification of frequently occurring types of queries that merit further attention. Two categories of keywords were found to be used most often: names of people and names of television programmes. The top 50 most used queries are depicted in Figure 1.

Further analysis of the user logs will link the frequency information to information about the success of a search. Success can be measured by the amount of times an item in the result list is viewed or ordered. By identifying which types of searches give the user a satisfying result, and which do not, we can target our research towards those types that need improvement. MuNCH has set out to enrich the log files to enable such detailed analyses.

In the next section, we will outline an architecture for a television broadcast archive that allows us to answer these different kinds of search goals. video, in combination with previous annotations by humans performed on the video, are presented to the user of the system by the interface.

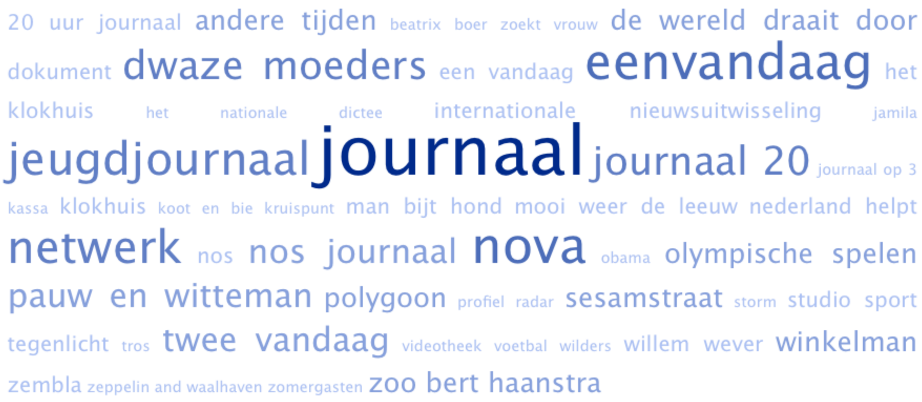

Figure 1: Tag cloud of top 50 most used search queries, measured over a period of 10 months in 2008.

3 Components of a Video Retrieval System

There is no single retrieval technique that will accommodate the diverse user needs sketched above, rather, multiple techniques need to be used together. As a consequence, systems for processing and retrieval of audio-visual archives will always be the sum of a large number of components. In this section we aim to give the reader an overview of how these different techniques will work together as components in an integrated retrieval system.

The MediaMill semantic video search engine (Worring et al. 2007) is such an integrated system. Many of MuNCH's contributions are – or will be – implemented as components in the MediaMill system. To give an idea of the complexity and structure of such a system, Figure 2 depicts the envisioned architecture of the MediaMill system. For an exhaustive description of the architecture, we refer to Worring et al. (2007). Here, we give a summary of the most relevant components for our purposes in order to demonstrate how MuNCH's contributions fit within the bigger picture of a working video processing and retrieval system.

To tackle the diverse user needs in video retrieval, we process a large variety of resources: the audiovisual content of the broadcast material, descriptions of the material by cataloguers, background knowledge about the domain of Dutch television in a thesaurus, and textual information in or about the video. In addition, we use transcriptions of the speech in the television material, provided by a speech recognition component . In order to process the audio-visual content of the broadcast material, it is first split into an visual stream and an audio signal. The visual stream is segmented into shots, where a shot is a cohesive piece of video recorded in one take. At the same time, important stills or "keyframes" are selected that characterise the shot. The audio signal is segmented into audio shots that do not necessarily have the same length as the image shots. Low-level features are extracted from the shots; motion features in the visual shots, still

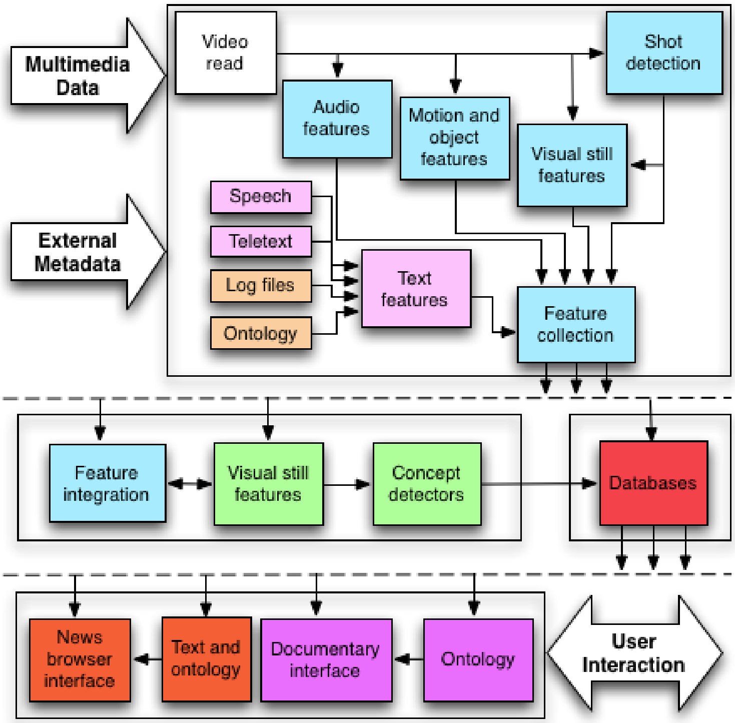

Figure 2: Envisaged architecture of a processing system for digital video repositories.

features (e.g., colour and texture features) in the keyframes and audio features (e.g., audio events and speech) in the audio shots.

All text associated with the video is analysed: subtitles, speech transcriptions and externally sourced descriptions from electronic program guides and websites. Text features are extracted and indexes are built to enable fast search. The low-level features are integrated and used to infer the presence of high-level concepts in a shot. First, concept detectors are trained with positive and negative examples. Once a detector is trained it is used to automatically annotate shots with concepts. The annotations are not absolute but rather an estimation of the likelihood that the concept is present in the shot. In the next section we will further detail the creation and use of concept detectors.

In the user interface of the retrieval system, a user will be able to search the archive of television broadcasts using not only the manual annotations, but also the (estimated) concepts that are detected in the shots, the speech and the external texts. Background knowledge from ontologies and thesauri is used to broaden a user query, and to form a bridge between the user queries and the detected concepts. The next section elaborates on our research using the thesaurus of Sound and Vision.

4 Automatic Methods for Improving Access to Television Broadcast Archives

Sound and Vision employs a team of professional cataloguers to create descriptions of its broadcasts. They record metadata such as titles, dates, makers, copyrights and carriers. Depending on the perceived importance of a broadcast, and on available resource, they may also provide textual descriptions of what can be seen in the programme, of its subject matter, and assign terms from a thesaurus to it. These catalogue entries are used as input for a text-based retrieval engine called iMMix, which allows users access to the archives.

In section 2 we have identified three categories of user queries. Let us examine the extent to which the catalogue information supports these types. *Known item queries* are issued by users who know exactly which broadcast they wish to find. The catalogue entry for a programme almost always contains basic production data such as the broadcast name, medium, and copyright owner, so the manually created entries should suffice in this case. *Subject queries* are harder, since there may be multiple ways to express the same subject, only one of which is used by the annotator. Queries for *sequences, shots and quotes* can only be answered for programmes that have been annotated on the level of individual shots, which is only a small portion of the entire archive.

MuNCH has investigated various techniques to improve these last two query types, and we will present them below. All techniques, with the exception of the concept selection methods that we describe, were evaluated on Sound and Vision data. This was possible because in 2007 and 2008, Sound and Vision provided 400 hours of video material to the TREC Video Retrieval Evaluation

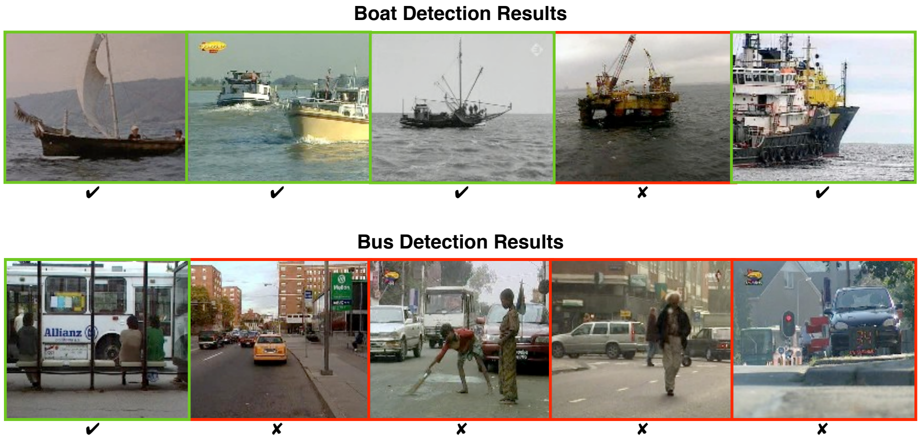

Figure 3: Results of semantic concept detection on video data from the Netherlands Institute of Sound and Vision. Correct detection is indicated by a check; incorrect detection by a cross.

(TRECVID) benchmark. Research groups from all over the world participate in this benchmark to test their content-based retrieval systems. a yearly conference, the participating research groups come together to compare their results. The data from Sound and Vision combined with the queries and ground truth from the TRECVID benchmark form a perfect testbed for large scale experiments. There is a further advantage: researchers from around the world are now working on the disclosure of Sound and Vision data.

4.1 Detecting Concepts in Video

Advances in the field of multimedia analysis are now opening up possibilities for automatic annotation. One technology that has high potential for providing access to shots and sequences in broadcast material is that of *concept detection*. A concept detector is an algorithm that identifies the likelihood of a certain object or scene being present in a piece of video, based on a machine-learned representation of the object or scene. An example of the output of two concept detectors built for the Sound and Vision collection is given in Figure 3. In this section we will give a summary of the approach we use to construct concept detectors for the Netherlands Institute of Sound and Vision, which is described in greater detail in Snoek et al. (2008).

First we extract low-level visual features from keyframes representing shots in the video collection. These visual features describe basic characteristics of the videos, such as colour distribution, textures, and edges. From this collection of low-level features we deduce the presence of a number of simple "proto-concepts" that describe the composition of a shot, such as *vegetation*, *water*, and *sky*.

Now that the videos have been described in terms of their visual characteristics, we are ready to develop concept detectors. To create a concept detector for a given concept, we obtain a collection of shots as training examples. These shots are labelled with respect to the presence or absence of a particular concept. A Support Vector Machine classifier (Vapnik 2000) is trained on these examples to construct a model of the high-level concept. Given a new shot (or more precisely, the features and proto-concepts of its keyframe), the classifier will be used to predict the likelihood that the semantic concept is present. We call such a classifier a concept detector.

Given sufficient training material, classifiers can be trained to detect the presence of various concepts such as *car*, *tree*, *person*, *snow* and *animal*. We currently have a growing lexicon of approximately 500 concept detectors in the domain of news broadcasts and documentaries. As can be seen in Figure 3, concept detectors vary in quality, and are rarely perfect. Some deciding factors in the quality of the detectors are the number of examples (both positive and negative) and the extent to which the characteristics of a concept can be described by our 15 proto-concepts. Because we use supervised learning we need a diverse range of examples for the classifier to be able to generalise from them. For example, if all examples of the concept *cityscape* are from commercial centres with skyscrapers, it is unlikely that a shot that was taken in a historic city centre will be correctly classified as cityscape. Although more research is needed to improve the concept detectors, they have the potential to provide detailed, shot-level annotations of large quantities of data. In the context of Sound and Vision this could mean better answers to queries for sequences, shots and quotes.

4.2 Concept Selection

Once a vocabulary of hundreds of concept detectors has been created, the user is able to exploit them for search. Selection of a (visual) concept detector appropriate to the query can allow users to quickly retrieve a list of relevant video fragments. With a large vocabulary it requires significant effort to manually select the best concept detector to use for a query, as we found in a focus group study aimed at this task (Huurnink, Hofmann, and de Rijke 2008). Therefore we have investigated how to automatically select the best concept detector for a particular search query. Here we used a combination of all three research fields of multimedia analysis, language technology, and semantic technologies, as applied to the use case of broadcast news retrieval.

Based on the different forms of query input that might be available — namely query text, and perhaps some example videos — we identified three different approaches for selecting the most concept appropriate detector for a particular user query:

Text Matching Associate each concept detector with a textual description. At retrieval time, select the concept detector with the best match between the concept detector description and the query text.

Ontology Querying Associate each concept detector with terms in a structured thesaurus. At retrieval, automatically link the query text to terms

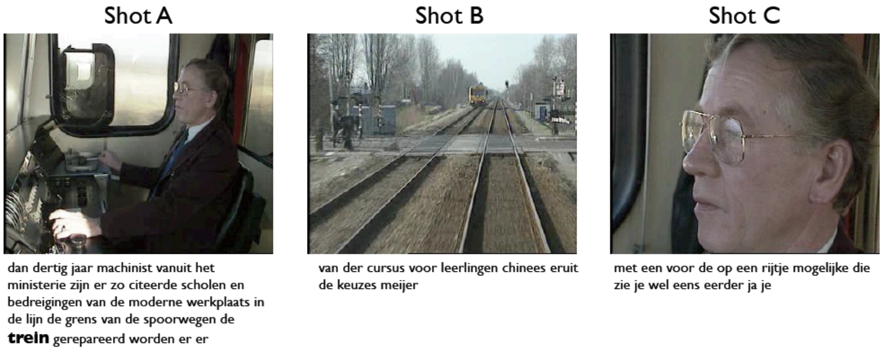
Shot A
Shot B
Shot C
dan dertig jaar machinist vanuit het ministerie zijn er zo citeerde scholen en bedreigingen van de moderne werkplaats in de lijn de grens van de spoorwegen de **trein** gerepareerd worden er er
van der cursus voor leerlingen chinees eruit de keuzes meijer
met een voor de op een rijtje mogelijke die zie je wel eens eerder ja je

Figure 4: An illustration of the temporal displacement between the mention of a train in the speech transcript, and its occurrence in video.

in the structured thesaurus and use ontology reasoning to select the closest detector to the query.

Semantic Visual Querying Build a visual model for each concept detector. At retrieval time, score the example videos in the query according to each concept detector, and select the detector with the closest visual match.

We evaluated these three concept selection methods with a search task. While the three strategies showed similar performance over a set of 24 search queries, we found that the text matching and ontology querying methods performed well for one set of queries, while semantic visual querying performed well for a different set of topics. This implies that when implementing concept selection, public broadcast archives would do well to consider incorporating multiple complementary methods originating from different fields.

4.3 Utilising Speech Transcripts for Search

Advances in speech recognition technology have made it possible to automatically transcribe spoken words from audio sound tracks. Though prone to errors, speech transcripts can still prove an important source of information for search. In MuNCH, we considered the specific example of language technology as applied to searching through automatic speech transcripts, specifically for the task of finding objects in the visual channel.

Video is characterised by a *temporal mismatch* between the mention of an object in the speech channel and its occurrence in the visual channel. These two sources of information are not necessarily synchronised. An example is given in Figure 4. Dutch readers will notice that the quality of the speech transcriptions is not perfect. Nevertheless, the speech transcripts can still contain valuable clues for search. For example, imagine that we need to find stock shots of trains. The Dutch word for train, *trein*, has been recognised in the speech track of shot A. A train can be seen in the visual track of the next shot, shot B. So when we are looking for trains, the occurrence of the word *trein* indicates that we are likely to find a train in some of the shots that surround it. Moreover, the closer a shot is to the word, the more likely that shot is to contain a train. This property can be used to advantage when searching through large video databases with speech transcripts. By analysing the characteristics of displacement, we discovered systematic shifts between spoken words describing objects, and the visual occurrence of those objects (Huurnink and de Rijke 2007). We performed such an analysis on a corpus of news video, and found that objects tend to be seen in the shot *after* they are mentioned in this kind of material. Incorporating this property into a retrieval model improves search performance for visual queries.

Similarly to search with concept detectors, speech-based search can provide users of the Institute's archive focused access to the content of video. Where concept-based search allows search on a restricted vocabulary of objects and scenes that can be seen in a broadcast, speech-based search allows users to search on dialogue, which may not only be indicative of what can be seen, but also of content in terms of topics and ideas. We expect that these two types of search will complement each other when implemented in the context of the public television broadcast archive.

4.4 Enriching Catalogue Descriptions

While in the previous subsections we have presented methods to create access to the archives of Sound and Vision without the need for manual annotation, we will now focus on a technique to make the most of the existing annotations that are made with the Sound and Vision thesaurus, the GTAA. The GTAA is an in-house thesaurus that has been designed specifically for cataloguers to use when manually describing video data. A more detailed description of the GTAA can be found in (Gazendam et al. 2009) in this journal.

Search using the manual annotations is very precise and rarely results in incorrect results - it has a *high precision*. However, because manual description of video data is so time-consuming, the cataloguers are instructed to only focus on the main topic and use a small set of terms to describe a programme. A low number of annotations per programme can lead to few results being returned in the search process - a *low recall* problem. To alleviate this problem we have put to use the structure of the thesaurus: queries are expanded to also include closely related thesaurus terms (Voorhees 1994), and thus return a broader set of results and a higher recall. For example, a query for Weather will be broadened by terms such as Climate, Rain and Wind, since the GTAA contains links between these terms and the term Weather. We expect that especially *subject queries* will benefit, since the question what a programme is about can be answered in different ways and at different levels of specificity (i.e., with different GTAA terms). Differences between the terms used by a a searcher and a cataloguer can be alleviated to some extent by including a range of related thesaurus terms.

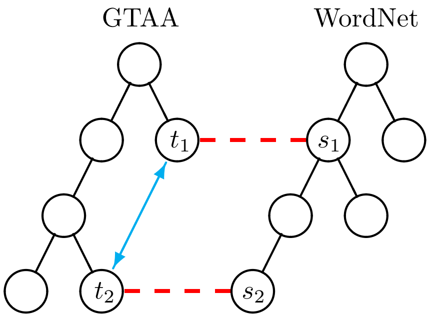

Figure 5: Using the mapping to WordNet to infer relations within the GTAA. Mappings between GTAA and WordNet are shows as red dashed lines, the newly inferred relation as a blue arrowed line.

This type of query expansion relies on a rich thesaurus structure with many interrelated terms. In contrast, the GTAA is a typical local thesaurus that is limited in breadth and depth. In close cooperation with the CHOICE project[1], we have therefore enriched the GTAA thesaurus structure by taking advantage of the rich semantic knowledge of an external resource, the English-American WordNet (Fellbaum 1998). First, we map terms in the GTAA thesaurus to concepts (*synsets*) in WordNet. Second, based on this mapping, we enrich the structure of the in-house GTAA thesaurus by inferring new relations between terms within the thesaurus (Hollink, Malaisé, and Schreiber 2008).

Figure 5 illustrates how a relation between two terms in the GTAA is inferred from their correspondence to WordNet synsets. If GTAA term t_1 corresponds to WordNet synset s_1 and GTAA term t_2 corresponds to WordNet synset s_2, and the two synsets s_1 and s_2 are closely related, we infer that the two GTAA terms t_1 and t_2 are also related. The two WordNet synsets are considered to be 'closely related' if they are connected though either a direct relation without any intermediate synsets or an indirect relation with one or two intermediate synsets. The latter situation is shown in Figure 5. From all WordNet relations, we used only part-of and subclass relations, and their inverses whole-of and superclass. A previous study demonstrated that other types of WordNet relations do not improve retrieval results when used for query expansion (Hollink, Schreiber, and Wielinga 2007).

A total of 1039 pairs of GTAA terms was newly related. Examples are:

`squid`	-	`colouring (pigment)`
`pharmacy`	-	`medicine`
`barbecues`	-	`picnics`
`national anthems`	-	`music`
`pearls`	-	`jewelery`
`fjords`	-	`seas`

[1] http://ems01.mpi.nl/CHOICE/ (15/03/09)

The aim is to use these newly inferred relations to expand queries in the same way that existing thesaurus relations are used. To investigate if the additional structure does indeed improve retrieval results, we performed an experiment comparing search with the original in-house thesaurus to search with the enriched thesaurus. Retrieval using only the inferred structure performed comparably to retrieval using only the original structure. This suggests that it is possible to use an external resource to enrich an otherwise unstructured vocabulary and base retrieval on the inferred relations. For example, we see possibilities to enrich lexicons of concept detectors such as those described previously in this paper.

Using the inferred structure in combination with the original structure improved search performance moderately. A larger set of newly inferred relations is needed to confirm that it is beneficial to enrich an already structured thesaurus. Generally speaking, the results are a promising example of how library science can be combined with recent semantic web techniques such as thesaurus mapping and semantic query expansion.

5 Discussion and Conclusions

In the MuNCH project an exchange of ideas took place, not only between its three academic partners but also between the academic and the cultural heritage worlds. This has led to a growing awareness in the participants as to how much can be gained by looking for solutions across the boundaries of separate disciplines. Much of this awareness stems from an initial analysis of the needs of users searching through the broadcast archives at Sound and Vision. Although this analysis is still in an early stage, a preliminary categorisation of query types and prioritisation of frequently posed queries has motivated the investigation of different approaches to video retrieval. A new logging module designed in cooperation with MuNCH will provide more detailed information on user behaviour. This information will be used by the Institute to fine-tune its services, and in addition we hope it will prove a rich source of information for the scientific community.

In this paper we have presented several fruitful cooperations between Sound and Vision and the three scientific disciplines brought together in MuNCH: multimedia analysis, language technology, and semantic technologies. One concrete example of such cooperation is interdisciplinary investigation into how to identify the best concept detector to answer a given information need. A second example is the use of semantic techniques to enrich the current thesaurus and use of the catalogue, showing that semantic web techniques and library science work well together.

Considering the unique combination of scientific disciplines present in MuNCH, these complement each other well in the context unlocking the archives at Sound and Vision. They can all contribute different sources of information for users of the archive to search on: with multimedia analysis we can create concept detectors that automatically find objects in video shots; through language technology we can make optimal use of textual information such as automatic speech transcriptions; and by applying semantic technologies we can enrich the Institute's catalogue of thesaurus descriptions with inferred relationships. These contributions provide Sound and Vision with insight into the potential of automatic techniques for augmenting the existing search possibilities provided by the catalogue descriptions, especially for those users who are interested in searching for specific video sequences, shots, and quotes. The prospect of incorporating multiple types of evidence for search is promising in that it can potentially support a variety of search tasks.

Although MuNCH has produced promising results with its combinations of research fields, there is still more to do. MuNCH plans research into incorporating the multiple channels of manually and automatically derived information into a single system. Another important direction for future work is scaling multimedia analysis techniques to the speed and efficiency required by an institute such as Sound and Vision, as such techniques currently demand a prohibitive amount of processing power. We expect that this problem is solvable with advances in computing power and through the creation of efficient algorithms.

From the perspective of Sound and Vision, a number of future research directions have become apparent. First of all, there is increasing interest in peer-to-peer systems and Web 2.0 technologies in the information retrieval community. These technologies provide an infrastructure that allows for filtering, recommending and tagging video content in a novel way (Cho and Tomkins 2007). In the context of Sound and Vision, we think of the use of preferences of "buddies", recommendations based on similar profiles and investigating the context (i.e., blogs, websites) in which an item is used and the comments and tags others have provided. Its potential for use in search systems has been recognised but more evaluation is required, especially in relation to existing annotation technologies.

A second research area for the near future is the analysis of people's search behaviour, for example using Information Foraging Theory. Evaluations in the domain of retrieval provide evidence that browsing is an important mode for users examining (video) collections (Savolainen and Kari 2006; Van Houten 2009). The Sound and Vision video retrieval system should support this type of interaction, based on available metadata, visual features and using vocabularies; technologies used in MuNCH. A future system that integrates these techniques will have an increased complexity, providing options for several search mechanisms, and allowing for searches on programme level, but also within a given programme. This underlying complexity, however, should not be visible on the graphical user interface that end-users will interact with and should match their cognitive skills (Norman 1998). Therefore, future work will need to address Human-Computer Interaction issues.

Concluding, MuNCH has shown the Netherlands Institute of Sound and Vision the potential of automatic techniques to help unlock its content in the near and far future. New uses of their video data, user data and thesaurus have created an awareness of the enormous possibilities of these valuable resources for the academic as well as the cultural heritage fields. The user needs at Sound and Vision have demonstrated to the project's academic partners that no single discipline suffices and that research is needed not only in combining technologies, but also in methods to find the right combination for each particular query.

Acknowledgements

The MuNCH project is supported by the Netherlands Organisation for Scientific Research (NWO) programme for Continuous Access To Cultural Heritage (CATCH) under project number 640.002.501. This research was also supported by the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments (http://www.stevin-tst.org) under project number STE-09-12, and by the NWO under project numbers 017.001.190, 640.002.501, 612.066.512, 612.061.814, 612.061.815, 640.004.802.

References

Cho, Junghoo, and Andrew Tomkins. 2007. "Guest Editors' Introduction: Social Media and Search." *IEEE Internet Computing* 11 (6): 13–15.

Edmondson, Ray. 2004. *Audiovisual Archiving: Philosophy and Principles.* Paris, France: UNESCO.

Fellbaum, C., ed. 1998. *WordNet: an electronic lexical database.* Cambridge, MA, USA: MIT press.

Gazendam, Luit, Véronique Malaisé, Annemieke de Jong, Christian Wartena, Hennie Brugman, and A. Th. Schreiber Schreiber. 2009. "Automatic Annotation Suggestions for Audiovisual Archives: Evaluation Aspects." *Interdisciplinary Science Reviews, special issue on Continuous Access To Cultural Heritage.*

Hollink, Laura, Véronique Malaisé, and Guus Schreiber. 2008, December. "Enriching a Thesaurus to Improve Retrieval of Audiovisual Documents." *the Third International Conference on Semantic and Digital Media Technologies (SAMT).* 47–60.

Hollink, Laura, Guus Schreiber, and Bob Wielinga. 2007. "Patterns of Semantic Relations to Improve Image Content Search." *Journal of Web Semantics* 5:195–203.

Huurnink, B., and M. de Rijke. 2007, September. "Exploiting Redundancy in Cross-Channel Video Retrieval." *Proceedings of the 9th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR 2007).* ACM Press, 177–186.

Huurnink, Bouke, Katja Hofmann, and Maarten de Rijke. 2008, October. "Assessing Concept Selection for Video Retrieval." *Proceedings of the 10th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR 2008).*

Norman, Donald A. 1998, August. *The Design of Everyday Things.* Cambridge, MA: MIT Press.

Savolainen, Reijo, and Jarkko Kari. 2006. "Facing and bridging gaps in Web searching." *Information Processing & Management* 42 (2): 519 – 537.

Smeulders, Arnold W. M., Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. 2000. "Content-Based Image Retrieval at the End of the Early Years." *IEEE Trans. Pattern Anal. Mach. Intell.* 22 (12): 1349–1380.

Snoek, Cees G. M., Koen E. A. van de Sande, Ork de Rooij, Bouke Huurnink, Jan C. van Gemert, Jasper R. R. Uijlings, J. He, Xirong Li, Ivo Everts, Vladimir Nedović, Michiel van Liempt, Richard van Balen, and Fei Yan an. 2008, November. "The MediaMill TRECVID 2008 Semantic Video Search Engine." *Proceedings of the 6th TRECVID Workshop.* Gaithersburg, USA.

Van Houten, Yntze. 2009. "Searching for videos: The structure of video interaction in the framework of information foraging theory." Ph.D. diss., Telematica Instituut Enschede.

Vapnik, V. N. 2000. *The Nature of Statistical Learning Theory.* 2nd. New York, USA: Springer-Verlag.

Voorhees, E. 1994. "Query expansion using lexical-semantic relations." Edited by W. B. Croft and C. J. Rijsbergen, van, *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* New York, NY, USA: Springer-Verlag, 61–69.

Worring, Marcel, Cees G. M. Snoek, Ork de Rooij, Giang P. Nguyen, and Arnold W. M. Smeulders. 2007. "The MediaMill semantic video search engine." *Proc of IEEE Int Conf on Acoustics, Speech, and Signal Processing, ICASS07.* IEEE-Press.