Today's and Tomorrow's Retrieval Practice in the Audiovisual Archive

Bouke Huurnink
bhuurnink@uva.nl

Cees G.M. Snoek
cgmsnoek@uva.nl

Maarten de Rijke
derijke@uva.nl

Arnold W.M. Smeulders
arnoldsmeulders@uva.nl

ISLA, University of Amsterdam
Science Park 107, Amsterdam, The Netherlands

ABSTRACT

Content-based video retrieval is maturing to the point where it can be used in real-world retrieval practices. One such practice is the audiovisual archive, whose users increasingly require fine-grained access to broadcast television content. We investigate to what extent content-based video retrieval methods can improve search in the audiovisual archive. In particular, we propose an evaluation methodology tailored to the specific needs and circumstances of the audiovisual archive, which are typically missed by existing evaluation initiatives. We utilize logged searches and content purchases from an existing audiovisual archive to create realistic query sets and relevance judgments. To reflect the retrieval practice of both the archive and the video retrieval community as closely as possible, our experiments with three video search engines incorporate archive-created catalog entries as well as state-of-the-art multimedia content analysis results. We find that incorporating content-based video retrieval into the archive's practice results in significant performance increases for shot retrieval and for retrieving entire television programs. Our experiments also indicate that individual content-based retrieval methods yield approximately equal performance gains. We conclude that the time has come for audiovisual archives to start accommodating content-based video retrieval methods into their daily practice.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*

General Terms

Performance, Experimentation, Human Factors, Measurement

1. INTRODUCTION

Progress in digital recording, storage, and networking technology has enabled large-scale ingestion and dissemination of multimedia material. As a consequence, audiovisual archives responsible for guarding and saving 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 [5]. Users of the archive search on these textual descriptions and receive (rankings of) complete television programs as results. However, more and more user groups require and demand access to video *fragments* rather than entire programs — video fragments accounted for 66% of purchases in one recent study of a broadcast archive [10]. 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 [3].

Our central aim in this paper is the following:

> *To explore the potential of content-based video retrieval for enhancing the retrieval practice in the audiovisual archive of today and tomorrow.*

Existing evaluation initiatives are unsuited to address the aim of investigating content-based video retrieval in a real-world setting, as their queries are not based on real-world queries, and no manually created metadata (which is often present in the real world) is included in the experiments. Therefore, we propose an evaluation methodology tailored to the specific needs and circumstances of the audiovisual archive, directed by four research questions:[1]

RQ1 What is the potential of content retrieval to answer today's queries in the archive, and queries as they might be formulated in the archive of the future?

RQ2 What can content retrieval add to search performance when combined with current archive search capabilities?

RQ3 Can content retrieval help those users that wish to retrieve entire programs?

RQ4 Which content retrieval methods should be given priority for integration into the archive?

Ultimately, our answers to these questions benefit policy makers at audiovisual archives who are facing the limitations of today's manual annotation practices and are considering incorporating content

---

[1]For ease of reading, we will refer to content-based video retrieval as *content retrieval* from this point onward.

retrieval into their work-flow. In addition, our answers are of interest to researchers as they apply content retrieval outside of the usual laboratory benchmark setting.

Our evaluation methodology integrates multiple perspectives in order to answer the research questions. First, we define three query sets, including both current user queries and those that might be issued in a future content retrieval-equipped archive. Second, we build a search engine that exploits both manually created and automatically generated annotations. Third, 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 retrieval might be integrated into tomorrow's audiovisual archive.

The contributions of this quantitative study of how content retrieval can help improve retrieval in the audiovisual archive are four-fold.

- First, we present an experimental methodology for assessing the potential of content retrieval for audiovisual archives.

- Second, we present a method for combining manually created program annotations with automatically generated shot annotations plus insights into its effectiveness.

- Third, we present a method for using automatically generated shot annotations to retrieve entire programs, with insights into its effectiveness.

- Finally, we contribute 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.[2]

The rest of this paper is structured as follows. We discuss related work in Section 2. We present our evaluation methodology in Section 3. In Section 4 we outline our experimental setup. Results are presented in Section 5. We end this paper with conclusions and recommendations for the archive in Section 6.

2. RELATED WORK

We first review trends in audiovisual retrieval from the content perspective, followed by a summary of crossover studies that incorporate the practitioner's perspective from the audiovisual archive.

2.1 The content perspective

The literature on content-based video retrieval and its evaluation is vast and impossible to cover here completely [25]. Instead, we identify three dominant content retrieval methods according to the type of video search input. *Transcript-based search* utilizes automatic speech recognition transcripts and machine translation of spoken dialog to retrieve video fragments given a textual query. While originally proposed over a decade ago [2, 33], the method is still very relevant today [9, 40, 41] especially when high-quality speech recordings are available. Transcript-based search provides indirect access to visual content, relying on the mention of visible objects and scenes in the video dialog *Feature-based search* allows direct access to visual information by representing keyframes in terms of low-level visual descriptors, which are then matched to query images [24]. This search method has evolved from exploiting basic similarity metrics between global image histograms of video fragments, to more advanced methods incorporating invariant keypoint descriptors [11, 30] and online learning [15, 17]. While this method can give accurate results, especially when provided with distinctive examples, it is difficult for humans to interpret. The third

content retrieval method is *Detector-based search*. This method utilizes shot-based detection scores for a given human-defined concept — such as a horse, a telephone, or a musical instrument — to retrieve video fragments. Similar to feature-based search the state-of-the-art is based on invariant keypoint descriptors [11, 30], which are softly assigned to a stacked codebook [31], and combined with kernel-based machine learning [27]. To cater for retrieval the detectors need to be selected and combined with the aid of query analysis using text, ontology, or visual matching [6, 18, 27, 34].

For improved retrieval performance, results from the different content retrieval methods may be combined, e.g., [33]. Fusion of multimedia search results is strongly query-dependent, and an area of ongoing research [13, 36, 39]. Approaches include query-class dependent combination schemes [15, 39], reranking of an initial result list [14, 28], and query-adaptive approaches [13].

Content-based analysis and video retrieval methods have been evaluated extensively in TRECVID [23]. The aim of TRECVID is to promote progress in content-based analysis of and retrieval from digital video via open, metrics-based evaluation using a common data set. TRECVID has been of pivotal importance in assessing content retrieval methods on their relative merit. While valuable, TRECVID's search tasks are not without criticism [7, 25, 35]. For example, it has been found difficult to replicate search experiments. In addition, it has been argued, that search topics are overly complex, limited in number, and drifting away from a real-world video retrieval practice.

2.2 The practitioner perspective

With an increasing amount of digitization in the audiovisual archive, a number of crossover efforts have used archive data to aid content retrieval, or conversely have studied attitudes towards content retrieval methods in the archive. In the category of using archive data to aid retrieval, Tsikrika et al. [29] utilize logged user result clicks in a photographic archive to create training data for concept detection algorithms. Allauzen and Gauvain [1] use manually created metadata from an audiovisual archive to augment document-specific speech recognition. Zuurbier [43] performed a study of the attitude of a group of audiovisual archivists towards the use of content retrieval methods in their workflow, specifically towards using automatic speech transcription to annotate. He found that while archivists acknowledged the potential of such technology to help annotate video, they were sceptical about the usefulness of imperfect automatically generated annotations. One of the studies most closely related to this work is that of Carmichael et al. [3], who perform a user-based evaluation of a content retrieval system based on automatic speech transcripts in the audiovisual archive. They find that the system helps professional users interact with the archive retrieval system in a new way. Finally, the VideOlympics showcase [26] has evaluated the user side of content-based video retrieval systems. To the best of our knowledge no content retrieval evaluation methodology exists which is tailored to the specific needs and circumstances of the audiovisual archive.

3. EVALUATION METHODOLOGY

We use a quantitative system evaluation methodology to explore the potential of content retrieval for enhancing search performance in the audiovisual archive. System evaluation requires a collection of documents, a set of statements of information need (called "queries" in this paper), and relevance judgments indicating which documents in the collection should be returned for each query [32]. Existing evaluation initiatives utilize documents, queries, and relevance judgments that do not reflect retrieval practice in the archive. Therefore we develop an evaluation methodology that does. In par-
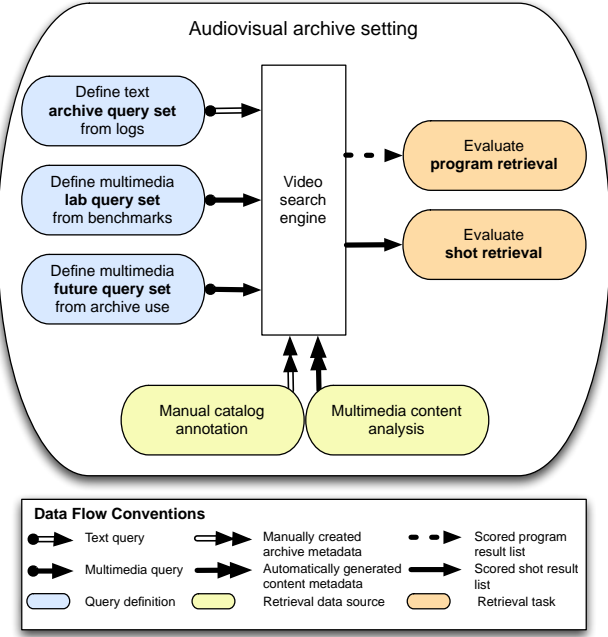
Figure 1: Evaluation methodology used to evaluate the potential impact of content-based video retrieval in the audiovisual archive. Note the inclusion of queries and retrieval data sources from the archive, as well as the archive-based program retrieval task.

ticular, we create (1) real-world queries derived from archive usage data and compare them to queries from common benchmark evaluations, (2) a video search engine based on manually created annotations from the archive; and (3) a program-level retrieval task, the current form of search in the archive. We summarize our methodology in Figure 1 and detail the individual ingredients next.

3.1 Audiovisual archive setting

Our study of content retrieval in the audiovisual archive takes place within the context of the Netherlands Institute for Sound and Vision, which we will refer to as "the archive." The Netherlands Institute for Sound and Vision is a good choice to represent "the audiovisual archive" for a number of reasons. It is growing rapidly, with (digital) television material being added to the archive as it is broadcast, and so far it has been impossible to manually annotate all of the new programs entering the archive. It represents a broader class of national broadcast archives, similar, for example, to the British BBC, the French INA, and the Italian RAI [38]. In addition, most of its users are searching for pieces of video to reuse in new television productions, and as such have a need to find fragments of video rather than consuming entire programs. Currently the archive caters for this need by allowing users to search for programs, which can then be browsed using a keyframe viewer or a video preview so that the desired fragment can be retrieved.

3.2 Query definitions

3.2.1 Query set 1: Archive queries

To create a set of Archive queries based directly on today's user needs, we make use of the archive's transaction logs. In other settings, searches and clicks from transaction logs have been used to create queries and relevance judgments for retrieval experiments [12, 21]. Our approach is different because we also include *purchase*

data, in addition to click data. We interpret a purchased video as a fulfilled information need, allowing us to consider the purchase data as relevance judgments in our evaluation [8].

We define an Archive query by first identifying all logged search sessions that resulted in a purchase from the archive's video collection. We then concatenate the text from the various searches in each session to form the final query. We exploit the purchase data as relevance judgements at the program-level. Relevant shots are identified within the start and end time of the purchased program. When an entire program is purchased, as in e.g., the second example in Table 1, we mark all shots within that program as relevant.

3.2.2 Query set 2: Lab queries

We create Lab queries that are representative of those used in content retrieval research by adopting them from several existing evaluation initiatives. Specifically, our Lab query set incorporates queries from the TRECVID 2007 and 2008 retrieval tasks [23], and the 2008 VideOlympics interactive retrieval showcase [26]. As the video collections used in these initiatives vary from year to year, the queries have relevance judgments on different collections. We performed additional relevance judging to identify relevant shots in the experimental collection used in this paper; a group of annotators manually labeled shots from the video collection as relevant or non-relevant using an interactive annotation tool [4]. Each annotator 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. Each annotator was able to browse through the video using transcript-based search, feature-based search, and detector-based search, as well as online learning, and associative browsing through the video timeline.

We use the relevance judgments at the shot level to create relevance judgments at the program level. We do so 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.

3.2.3 Query set 3: Future queries

Turning back to the needs of archive users, we create a set of Future queries. These are based on logged user needs, but reformulated in terms of an archive retrieval system that includes content retrieval capabilities. Today's logged archive queries and purchases are not necessarily well suited for evaluating content retrieval. Queries regularly do not contain words describing the required video content, consisting rather of program titles or technical codes [10]. Purchases do not always clearly delineate the video in terms of required visual content, for example when an entire program is purchased. 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. Once again, we create a set of queries using transaction logs. However, instead of directly utilizing logged searches, we reformulate them as multimedia queries.

To create the Future queries, we selected 24 logged user sessions that resulted in a purchase of audiovisual data. The information contained in the sessions included searches, result clicks, and purchases. An independent query creator from the archive was given the information from each session, and was asked to develop queries that she felt reflected the underlying information need of the broadcast professional. To be precise, the query creator was asked to: (1) scan the session to get an idea of the general information needs of the searcher; (2) view the video fragments that were or-

Table 1: Sample searches and purchases contained in the transaction log data from the audiovisual archive and 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.

User Searches	Purchase details	Randomly selected keyframes from purchase
shots f16 saab airplane shots	*Program title*: Zembla — The Defense Orders *Purchase duration*: 13s *Program duration*: 35m 32s	
noorderlicht *on* 1996-11-10	*Program title*: Noorderlicht — The Image of the Dolphin *Purchase duration*: 25m 13s *Program duration*: 25m 13s	

dered; (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. Once the query generation process was completed, two query selectors examined the information needs and selected those that were likely to have relevant examples in the experimental test collection. The text of each query was associated with 1–5 video examples so to turn it into a proper multimedia query. Relevant shots were identified in the same manner as for Lab queries.

3.3 Retrieval data sources

In today's archive, the main source of retrieval data used is a collection of manually created catalog entries that describe each program. We show an excerpt of such an entry in Figure 2. The archive structures its catalog entries using multiple information fields. In our evaluation methodology, we aggregate the different fields into three different types, namely: *free text*, natural language descriptions that describe and summarize the content of a program; *tags*, structured thesaurus terms that describe the people, locations, named entities, and subject areas that appear in or are the topic of a program; and *technical metadata*, technical information about a pro-

Field	Name
	Technical Metadata
Title	Noorderlicht — The Image of the Dolphin
Broadcast date	1996-11-10
Carrier number	HETBEELDVANDE-HRE000038DA.mxf
Carrier type	MXF
Carrier id	128646
	Free Text
Summary	Program with reports on scientific topics. In this episode, research by biologist Ken Marten on Ohau, one of the Hawaiian Islands, into the behavior of dolphins.
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[...]
	Tags
Genre	Educational; Magazine
Location	Hawaii
Person	<no entry>
Name	<no entry>
Subject	biology; dolphins; behavioral science; scientific research; intelligence; pain
Maker	Doornik, Jack van; Feijen, Joyce; Hattum, Rob van; Hermans, Babiche [...]

Figure 2: Excerpt from an example catalog entry from the audiovisual archive (translated into English). The catalog fields are divided into three different types: technical metadata, free text, and tags.

gram such as identification codes, copyright owners, available formats, and the program title. In addition to these manually created catalog entries, we utilize state-of-the-art multimedia analysis results produced by *transcript-based*, *feature-based*, and *detector-based* methods from the video retrieval literature, as detailed in Section 2. These three content retrieval methods together with the manually created catalog entries define the retrieval data sources for our evaluation methodology.

3.4 Video retrieval tasks

We consider two video retrieval task, organized by search unit.

3.4.1 Task 1: Shot retrieval

Users in the archive cannot currently retrieve shots, but over 66% of the orders in the archive contain requests for video fragments. Hence, shot-based video retrieval could allow these users to search through tomorrow's archive much more efficiently. Therefore, we include a shot retrieval task in our evaluation methodology. To adapt the program-level level catalog annotations for shot retrieval, we return the shots for each program in order of appearance.

3.4.2 Task 2: Program retrieval

Users in the archive currently retrieve entire programs, and tomorrow's archive is likely to continue support of this task. Therefore, we include a program retrieval task in our evaluation methodology. This requires an adjustment to the retrieval based on shot-based multimedia content analysis. To adapt the shot-level annotations for content retrieval, we employ an approach from the domain of *passage retrieval* [22]. We evaluated a number of approaches from the passage retrieval literature, and found the decay-based method [37] to work well in aggregating shot-level results for program retrieval.

4. EXPERIMENTAL SETUP

Now that we have outlined our evaluation methodology, we move on to describe the experimental setup. We summarize the statistics of our three query sets and their associated relevance judgments in Table 2. A visual overview of the future query set, which we created by analyzing visual information needs in the archive search logs, is given in Figure 3. As our video collection, we adopt the set of audiovisual broadcasts that the archive made available to the TRECVID benchmark in 2008. The test set of this video collection consists of over 100 hours of Dutch archived television broadcasts, 219 programs in total. The programs are diverse: the oldest pro-

Table 2: Statistics of the three query sets and their associated relevance judgments for shots and programs, which we created for evaluating video retrieval in the audiovisual archive.

Query set	Evaluation data		
	Queries	*Shots*	*Programs*
Archive	36	4,838	50
Lab	72	21,537	3,653
Future	29	4,007	485

gram was first broadcast in 1927, the most recent in 2004. The video collection has been pre-segmented [19] into 35,766 shots.

4.1 Video retrieval experiments

To answer our research questions related to the potential of content retrieval for enhancing the search practice in the audiovisual archive, we conduct the following three experiments:

- **Experiment 1:** *Shot retrieval with three video search engines using three query sets*

In this experiment, we simulate 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 RQ1 and RQ2.

- **Experiment 2:** *Program retrieval with three video search engines using three query sets*

In this experiment we simulate 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 RQ1, RQ2 and RQ3.

- **Experiment 3:** *Prioritizing content-based video search methods*

We examine the potential contribution of three different types of content-based search; namely transcript-based search, feature-based search, and detector-based search. This experiment aims at answering RQ4. We perform this experiment on the queries that are currently uncommon for the archive, namely the lab query set and the future query set.

Performance measure and significance tests For all three 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.

4.2 Video search engine implementations

Video search engine 1: catalog-based Our catalog-based 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 [16] the query terms. Retrieval is done using the language modeling paradigm [20]. To compensate for data sparseness and zero probability issues, we interpolate document and collection statistics using Jelinek-Mercer smoothing [42]. 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.
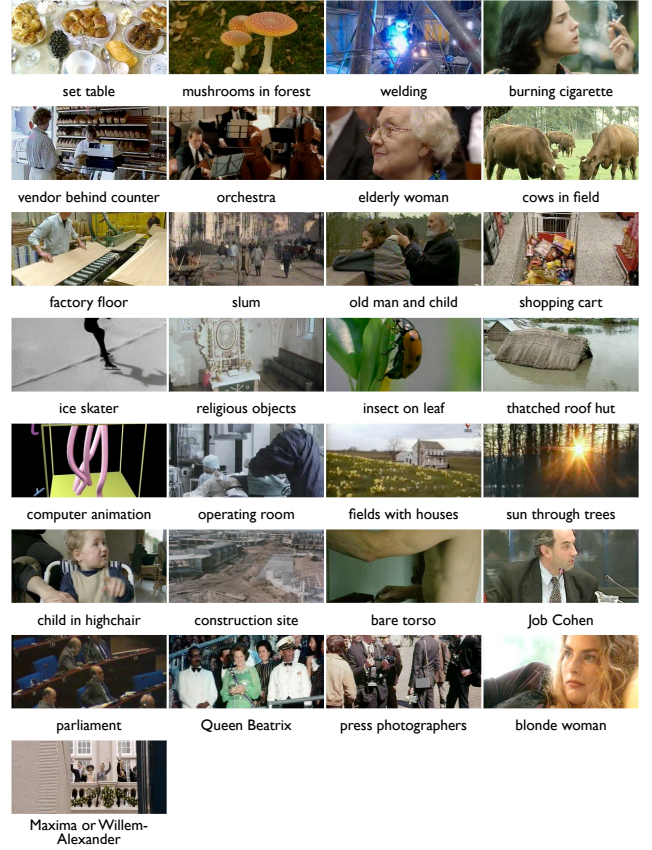

Figure 3: Visual overview of the *future query set*, which we derived by reviewing the logged behavior of users in the audiovisual archive.

Video search engine 2: content-based The content-based search engine is based on shot-based multimedia content analysis, covering transcript-based, feature-based, and detector-based search. We create a retrieval result for each of the three different types of search using the state-of-the-art methods described in [27]. 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).

Video search engine 3: future The future video 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 produce multiple search result that must be combined for a final retrieval outcome. 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. We perform fusion using the settings recommended by Wilkins [36], 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.

	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 3: 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, worse, or statistically indistinguishable from the score using the catalog-based video search engine.

5. RESULTS

We now move on to the results of our experiments. The numbers are summarized in Table 3. Additionally, Figure 4 highlights the different patterns in retrieval performance between query sets.

5.1 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 4a and Table 3 (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. The Future queries perform equally well using both of these search engines. The Future video search engine, which optimally combines the Catalog and Content engines, achieves significant improvements 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 good performance of the Catalog search engine to the nature of the judgment process. Recall that Archive queries and judgments are created by directly taking search and purchase information from the archive logs. When an entire program is purchased, all of the shots within the program are judged as relevant, and intra-video ordering does not make a difference. We leave for future examination with a larger data set the impact such factors have on the use of logged archive data to evaluate content retrieval.

In answer to **RQ1**, *What is the potential of content retrieval to answer the current queries in the archive, and queries as they might be formulated in the archive of the future?*, content retrieval alone is not enough to satisfy the needs of today's archive users. However, if future users state their information needs in content retrieval terms (as is the case for the Future queries) then both search engines perform equally well. We gain the most when combining content retrieval with retrieval using the catalog entries —which brings us to **RQ2**, *What can content retrieval add to search performance when combined with manual annotations from an archive?* Today's Archive queries, though less sensitive to content-based methods than other query sets, gain a significant performance increase by embedding content retrieval into today's practice. After combination, tomorrow's Future queries gain even more, with performance more than doubling.

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

Table 4: Performance in MAP for Experiment 3; shot retrieval for two (multimedia) query sets using three different content retrieval methods. ▲, ▼, and °, respectively indicate that a score is significantly better, worse, or statistically indistinguishable from the score of the remaining two content retrieval methods, from left to right.

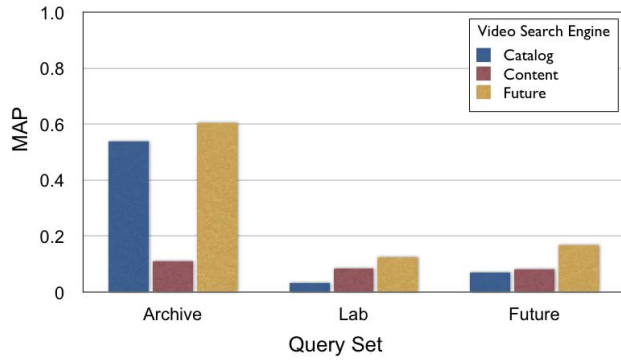## 5.2 Experiment 2: 3x3 program retrieval

The results of Experiment 2, i.e., program retrieval with three video search engines using three query sets, are given in Figure 4b and Table 3 (columns 5–7).

As was the case for shot retrieval, the Archive queries are much 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 much 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 Catalog search engine for these queries is due to the nature of the queries and judgments taken from the archive logs. The queries were taken directly from user searches, which were formulated in terms of the available archive catalog entries and contained technical metadata unsuited for content retrieval. The Lab and Future queries, on the other hand, perform better using the Content that 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 **RQ3**, *can content retrieval help those users that wish to retrieve 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 retrieval for program search: this is an artefact of the methodology used to create the queries and associated relevance judgments.

5.3 Experiment 3: prioritizing content search

The results for Experiment 3, i.e., shot retrieval with three different content retrieval methods, are shown in Table 4 and Figure 5. 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. These observations inform our answer to **RQ4**, *Which content 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 retrieval methods. Therefore we base our answer on other factors, namely: scalability, technological maturity, and ease of integration into the archive work-flow. The most suitable content retrieval method using these three criteria is transcript-based retrieval. Speech transcription has a relatively light processing footprint, has high accuracy for professionally recorded sound tracks, and can be queried using text alone.

(a) Experiment 1: 3x3 Shot Retrieval
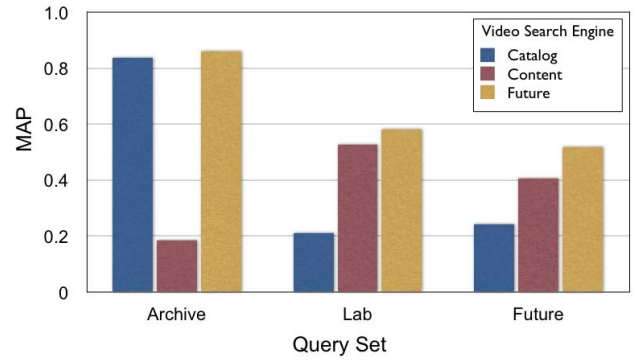
(b) Experiment 2: 3x3 Program Retrieval

Figure 4: Experimental results for shot and program retrieval in the audiovisual archive, across three query sets and three video search engines. Note the decrease in performance for the archive query set using content-based video search.
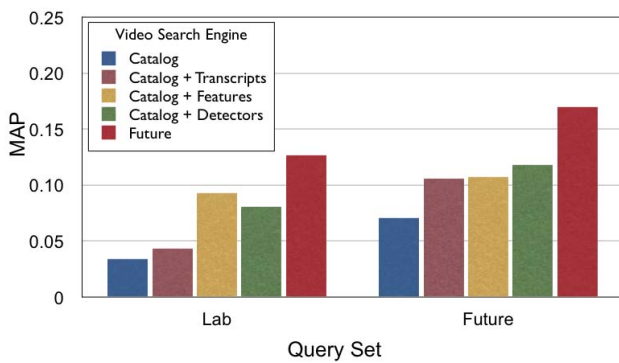

Figure 5: Shot retrieval MAP performance for the lab and future query sets, when combining retrieval results from the catalog-based video search engine with content-based video search methods.

6. CONCLUDING RECOMMENDATIONS

In this paper, we explored to what extent content-based video retrieval enhances today's and tomorrow's retrieval practice in the audiovisual archive using four research questions. To address these research questions, we proposed an evaluation methodology tailored to the specific needs and circumstances of the archive, including three query set definitions, three state-of-the-art content- and archive-based video search engines, and two challenging retrieval tasks that are grounded in a real-world audiovisual archive. Our first research question was, *what is the potential of content retrieval to answer the current queries of archive users, and queries as they might be formulated in the future archive?* We found that for today's (archive) queries, content retrieval was of limited use, but when archive logs were used to reformulate queries as they might be issued in tomorrow's archive, content retrieval outperforms traditional catalog-based video search engines of archives. To answer our second research question, *what can content retrieval add to search performance when combined with manual annotations from an archive?*, we found that a catalog-based video search engine supplemented with content retrieval yields performance gains up to 270%. Our experiments with program-level retrieval indicate a positive answer to the third research question, *can content retrieval help those users that wish to retrieve entire programs?*. We found that program retrieval with a content-based video search en-

gine improved upon catalog-based search by up to 147%. When we examined individual content retrieval methods, *Which content retrieval methods (transcripts, detectors, features) 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 approximately the same 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 retrieval into their work-flow. Due to issues of scale, technological maturity, and ease of integration into current retrieval capabilities, we recommend that audiovisual archives prioritize video retrieval using transcripts. Yet the biggest 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 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 retrieval and the audiovisual archive can mutually benefit from each other.

References

[1] A. Allauzen and J.-L. Gauvain. Open vocabulary ASR for audiovisual document indexation. In *ICASSP '05*, volume I, pages 1013–1016, 2005.

[2] 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 *ACM Multimedia '95*, San Francisco, USA, 1995.

[3] 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 *CBMI '08*, pages 93–100, 2008.

[4] O. de Rooij, C.G.M. Snoek, and M. Worring. Balancing thread based navigation for targeted video search. In *CIVR '08*, pages 485–494, New York, NY, USA, 2008. ACM.

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

[6] A. Hauptmann, R. Yan, and W.-H. Lin. How many high-level concepts will fill the semantic gap in news video retrieval? In *CIVR '07*, pages 627–634, New York, NY, USA, 2007. ACM.

[7] A.G. Hauptmann and W.-H. Lin. Assessing effectiveness in video retrieval. In *CIVR '05*, volume 3568 of *LNCS*, pages 215–225. Springer-Verlag, 2005.

[8] K. Hofmann, B. Huurnink, M. Bron, and M. de Rijke. Comparing click-through data to purchase decisions for retrieval evaluation. In *SIGIR '10*, 2010. To appear.

[9] B. Huurnink and M. de Rijke. Exploiting redundancy in cross-channel video retrieval. In *MIR '07*, pages 177–186. ACM Press, September 2007.

[10] 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. *J. American Society for Information Science and Technology*, 2010. To appear.

[11] Y.-G. Jiang, J. Yang, C.-W. Ngo, and A.G. Hauptmann. Representations of keypoint-based semantic concept detection: A comprehensive study. *IEEE Trans. Multimedia*, 12:42–53, 2010.

[12] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR '05*, pages 154–161, New York, NY, USA, 2005. ACM.

[13] L. Kennedy, S. Chang, and A. Natsev. Query-Adaptive Fusion for Multimodal Search. *IEEE*, 96(4):567–588, 2008.

[14] 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.

[15] 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 *TRECVID '08*, MD, USA, 2008.

[16] C. Monz and M. de Rijke. Shallow morphological analysis in monolingual information retrieval for Dutch, German, and Italian. In *CLEF '01*, pages 262–277, London, UK, 2002. Springer-Verlag.

[17] A.P. Natsev, M.R. Naphade, and J. Tešić. Learning the semantics of multimedia queries and concepts from a small number of examples. In *ACM Multimedia '05*, pages 598–607, Singapore, 2005.

[18] 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 *CIVR '06*, volume 4071 of *LNCS*, pages 143–152, Heidelberg, 2006. Springer-Verlag.

[19] C. Petersohn. Fraunhofer HHI at TRECVID 2004: Shot boundary detection system. In *TRECVID '04*, MD, USA, 2004.

[20] J.M. Ponte and W.B. Croft. A language modeling approach to information retrieval. In *SIGIR '98*, pages 275–281, NY, USA, 1998. ACM.

[21] F. Radlinski, M. Kurup, and T. Joachims. How does click-through data reflect retrieval quality? In *CIKM '08*, pages 43–52, 2008. ACM.

[22] G. Salton, J. Allan, and C. Buckley. Approaches to passage retrieval in full text information systems. In *SIGIR '93*, pages 49–58, New York, NY, USA, 1993. ACM.

[23] A.F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *MIR '06*, pages 321–330, 2006.

[24] J.R. Smith and S.-F. Chang. Visually searching the web for content. *IEEE MultiMedia*, 4(3):12–20, 1997.

[25] C.G.M. Snoek and M. Worring. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 4 (2):215–322, 2009.

[26] C.G.M. Snoek, M. Worring, O. de Rooij, K.E.A. van de Sande, R. Yan, and A.G. Hauptmann. Videolympics: Real-time evaluation of multimedia retrieval systems. *IEEE Multi-Media*, 15(1):86–91, 2008.

[27] 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. Nedovic, 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 *TRECVID '08*, MD, USA, 2008.

[28] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X.-S. Hua. Bayesian video search reranking. In *MM '08*, pages 131–140, New York, NY, USA, 2008. ACM.

[29] T. Tsikrika, C. Diou, A.P. de Vries, and A. Delopoulos. Image annotation using clickthrough data. In *CIVR '09*, New York, NY, USA, 2009. ACM.

[30] K.E.A. van de Sande, T. Gevers, and C.G.M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pattern Analysis and Machine Intell.*, 2010. In press.

[31] J.C. van Gemert, J.-M. Geusebroek, C.J. Veenman, and A.W.M. Smeulders. Kernel codebooks for scene categorization. In *ECCV '08*, Marseille, France, 2008.

[32] E.M. Voorhees. The philosophy of information retrieval evaluation. In *CLEF '01*, pages 355–370, London, UK, 2002. Springer-Verlag.

[33] 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.

[34] X.-Y. Wei, C.-W. Ngo, and Y.-G. Jiang. Selection of concept detectors for video search by ontology-enriched semantic spaces. *IEEE Trans. Multimedia*, 10(6):1085–1096, 2008.

[35] T. Westerveld. *Using Generative Probabilistic Models for Multimedia Retrieval*. PhD thesis, U. Twente, 2004.

[36] P. Wilkins. *An investigation into weighted data fusion for content-based multimedia information retrieval*. PhD thesis, Dublin City University, Dublin, Ireland, 2009.

[37] R. Wilkinson. Effective retrieval of structured documents. In *SIGIR '94*, pages 311–317, NY, USA, 1994. Springer-Verlag New York, Inc.

[38] R. Wright. Broadcast archives: Preserving the future. In *ICHIM '02*, pages 47–55, 2001.

[39] R. Yan, J. Yang, and A.G. Hauptmann. Learning query-class dependent weights in automatic video retrieval. In *MM'04*, pages 548–555, New York, NY, USA, 2004. ACM.

[40] J. Yang and A.G. Hauptmann. Exploring temporal consistency for video analysis and retrieval. In *MIR '06*, pages 33–42, New York, NY, USA, 2006. ACM.

[41] J. Yang, M. Chen, and A.G. Hauptmann. Finding person X: Correlating names with visual appearances. In *CIVR '04*, pages 270–278, New York, NY, USA, 2004. ACM.

[42] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.

[43] E. Zuurbier. *Onderzoek naar de haalbaarheid van Spoken Document Retrieval*. MSc. thesis, U. Twente, 2009.