

Content-Based Analysis Improves Audiovisual Archive Retrieval

Bouke Huurnink, Cees G. M. Snoek, *Senior Member, IEEE*, Maarten de Rijke, and Arnold W. M. Smeulders, *Member, IEEE*

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. In this paper, we take into account the information needs and retrieval data already present in the audiovisual archive, and demonstrate that retrieval performance can be significantly improved when content-based methods are applied to search. To the best of our knowledge, this is the first time that the practice of an audiovisual archive has been taken into account for quantitative retrieval evaluation. To arrive at our main result, 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, content purchases, session information, and simulators 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. A detailed query-level analysis indicates that individual content-based retrieval methods such as transcript-based retrieval and concept-based retrieval yield approximately equal performance gains. When combined, we find that content-based video retrieval incorporated into the archive's practice results in significant performance increases for shot retrieval and for retrieving entire television programs. The time has come for audiovisual archives to start accommodating content-based video retrieval methods into their daily practice.

Index Terms— Benchmark testing, content based retrieval, multimedia databases, search problems.

I. INTRODUCTION

As early as 1935, audiovisual archivists aimed to manually describe every shot in every video acquired by their archive [45]. However, due to the rapid increase in the number

Manuscript received August 12, 2011; revised November 25, 2011; accepted March 16, 2012. Date of publication April 05, 2012; date of current version July 13, 2012. This work was supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements nr 258191 (PROMISE Network of Excellence) and 288024 (LiMoSiNe project), IM-Pact BeeldCanon, STW SEARCHER, the Netherlands Organisation for Scientific Research (NWO) project nrs 612.061.814, 612.061.815, 640.004.802, 380-70-011, the Center for Creation, Content and Technology (CCCT), the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the WAHSP project funded by the CLARIN-nl program, the Dutch national program COMMIT under projects Infiniti and SEALINCmedia, and by the ESF Research Network Program ELIAS. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Shin'ichi Satoh.

B. Huurnink, C. G. M. Snoek, and M. de Rijke are with the Intelligent Systems Laboratory Amsterdam, University of Amsterdam, Amsterdam, The Netherlands (e-mail: bhuurnink@uva.nl; cgmsnoek@uva.nl; derijke@uva.nl).

A. W. M. Smeulders is with the Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands (e-mail: Arnold.Smeulders@cwi.nl).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TMM.2012.2193561

of videos coming into the archives, it soon became apparent that it was impossible to accomplish this goal through manual labor, given the limited human resources at their disposal. Archivists had to settle instead for describing whole videos, while occasionally providing more detailed within-video descriptions at the archivist's discretion. For example, titles and broadcast dates of videos are likely to be described, but individual shots and scenes are less likely to be described. Such an approach has limited searchers, especially those who do not know which specific video their desired footage appears in. However, 75 years later, the original dream of individually describing every shot in every video in an archive has now come within grasp. Where *manual* description of shots is not realistic, machines may fill the gap with the *automatic* shot descriptions associated with content-based video retrieval [41], [57]. Though these descriptions are not flawless, they may be helpful when searching through the archive.

Media professionals actively utilize audiovisual archives as a source for reusable material. The documentary maker requiring footage of Christmas trees from different cultures, and the news editor requiring footage of the Haiti earthquake for a news broadcast exactly one year after the disaster, can both turn to audiovisual archives to locate relevant footage themselves. In this task, they search through the archive using whatever annotations are available. Archives are struggling to reinvent themselves in the face of fully digital operations and growing user bases [32]. Yet, surprisingly, very little has been done to examine how content-based video retrieval will affect the searches of professionals searching in the audiovisual archive.

Our goal is to investigate how content-based video search can enhance the performance of traditional archive retrieval. We complement the old, manual, descriptions of the images in the archive with new, automatically generated, labels. Then, we measure the effect of combining them for queries typical of professionals searching an archive. Existing evaluation initiatives such as TRECVID [38], VideoCLEF [25], and Mediaeval [26] have been a valuable instigator in the advancement of techniques for content-based video retrieval. However, they are unsuited to assessing the potential impact of such techniques in a real-world setting such as the audiovisual archive. Their queries are not based on real-world queries, and generally no manually created metadata (which is often present in the real world) is included in the experiments. In contrast, our study is directed by three research questions:

RQ1: What is the potential of content-based video retrieval to answer the queries of professional searchers in today's archive, and their queries as they might be formulated in the future-world archive?

RQ2: To what degree can content-based video retrieval add to search performance when combined with current sources of archive search information?

RQ3: Can content-based video retrieval help those users that wish to retrieve entire programs?

Ultimately, the answers to our questions 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, as we include and investigate industrial searches and data sources (which have not been included in traditional benchmarks), the answers are also of interest to content-based video retrieval researchers.

The most important contribution of this paper, then, is a detailed investigation of how content-based video retrieval can improve audiovisual archive search. We develop an experimental methodology that allows us to quantitatively evaluate how retrieval performance for professional searches is affected. To enable replication of our experiments, we provide 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 We present methods for reconciling retrieval on information sources at different granularities (shot-level and program-level), and use weighted fusion to combine results. This allows us to provide an extensive analysis of the potential impact of content-based video retrieval in an archive.

The rest of this paper is structured as follows. We discuss related work in Section II. We present our evaluation methodology in Section III. In Section IV we outline our experimental setup. Results are presented in Section V, and are followed up with a query-level analysis in Section V-D. We end this paper with conclusions and recommendations for individual archives in Section VI.

II. 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.

A. Content Perspective

The literature on content-based video retrieval and its evaluation is vast and impossible to cover here completely [41], [57]. Instead, we identify three dominant content-based video retrieval methods according to the source of video retrieval data: *transcripts*, *detectors*, and *low-level features*. Together, these three sources have been extensively utilized in the content-based video retrieval community, as we will describe.

- 1) *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], [51], the method is still very relevant today [18], [58], [59] especially when high-quality speech recordings are available. This technology can gain high accuracy rates, with word accuracy

rates for high-quality recorded audio in well-defined domains such as broadcast news [7]. Accuracy is not as high for less well-defined domains and languages, where word error rates of 50%–60% are common [10]. However, due to redundancy, even transcripts with high error rates can still be effectively applied to retrieval tasks [11]. Transcript-based search provides indirect access to visual content, relying on the mention of visible objects and scenes in the video dialog.

- 2) *Low-level 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 [39]. 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 [22], [48] and online learning [28], [30]. While this method can give accurate results, especially when provided with distinctive examples, its reliance on basic compositional elements such as textures and edges makes them difficult for humans to interpret.
- 3) *Detector-based search*: 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 [22], [48], which are softly assigned to a stacked codebook [49], and combined with kernel-based machine learning [43]. Though detection can be noisy, progress in concept detection has been rapid, with the performance of concept detectors doubling over a period of three years [40]. To cater for retrieval, the detectors need to be selected and combined with the aid of query analysis using text, ontology, or visual matching [14], [31], [52]. Detector-based search allows access to shots directly on the basis of semantically interpretable visual content, but is limited by the number and quality of detectors available for search.

For improved retrieval performance, results from the different content retrieval methods may be combined, e.g., [51]. Fusion of multimedia search results is query-dependent, and an area of ongoing research [24], [44], [46], [54]. Besides merging results from different content retrieval methods, when retrieving whole programs, it is necessary to combine shot-level results from within the program. There has not been much work in this area, and approaches consist of assigning binary values to each shot and aggregating these to the program level [5], averaging the scores of all shots for a particular feature [6], using the maximum score of a shot in a program [37], or simply using the central shot in a program to represent the whole video [27]. Another approach is to use methods from *passage-based retrieval* [36], a research area which studies the problem of retrieving entire documents on the basis of their constituent passages [21].

Content-based analysis and video retrieval methods have been evaluated extensively in TRECVID [38]. 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

¹<http://ilps.science.uva.nl/resources/avarchive>.

importance in assessing content retrieval methods on their relative merit. While valuable, TRECVID's search tasks are not without criticism [13], [41], [53]. 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. Some attempts have been made to address this, for example by generating simulated queries based on the logs of real users [19].

In this paper we will unify the variety of different content-based video retrieval approaches by selecting a state-of-the-art method for each of transcript-based search, detector-based search, and feature-based search. We combine results from these methods using query-dependent weighted fusion, and incorporate sets of queries obtained from searches logged by an audiovisual archive.

B. 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-based video retrieval methods in the archive. In the category of using archive data to aid retrieval, Tsirikika *et al.* [47] 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. 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-based video 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 [42] has evaluated the user side of content-based video retrieval systems.

To the best of our knowledge, no content-based video retrieval evaluation methodology exists which is tailored to the specific needs and circumstances of the audiovisual archive, except for our previous work [21]. We extend this work in the following manner. In order to better answer how content-based video retrieval can be used to answer today's real-world queries, we added an extra set of 2190 simulated queries generated on the basis of the logged searches and purchases of professional users. These allow us to evaluate the impact at scale of content-based video retrieval methods on queries as they are currently issued in the archive, and also to evaluate the impact of content-based video retrieval when searching for entire programs. Furthermore, to gain insight into the effect of content-based video retrieval at the query level, we have performed a detailed query-level analysis of the effect of the envisioned Future search engine for both shot and program retrieval. Within our framework, we incorporate automatically generated metadata and state-of-the-art content-based video retrieval methods. In addition, we include queries from a real-world audiovisual archive, as well as a search engine based on manually created metadata. In this way we take both the content perspective and the practitioner perspective into account in our evaluation methodology.

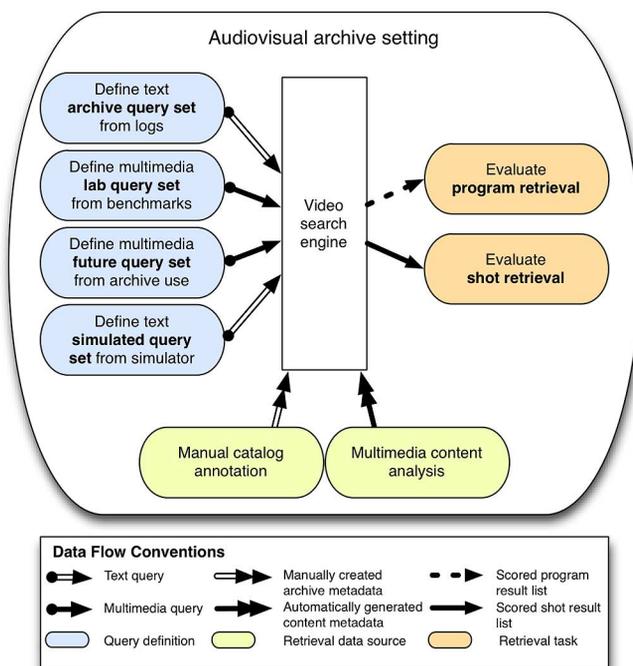


Fig. 1. Evaluation methodology used to evaluate the potential impact of content-based video retrieval in the audiovisual archive. Note the inclusion of queries and retrieval data sources from the archive, as well as the archive-based program retrieval task.

III. 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. 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 [50]. 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 particular, 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 Fig. 1 and detail the individual ingredients next.

A. Audiovisual Archive Setting

Our study of content-based video 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 [56]. In addition, most of

Field	Name
	<i>Technical Metadata</i>
Title	Noorderlicht — The Image of the Dolphin
Broadcast date	1996-11-10
Carrier number	HETBEELDVANDE-HRE000038DA.mxf
Carrier type	MXF
Carrier id	128646
	<i>Free Text</i>
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[...]
	<i>Tags</i>
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 [...]

Fig. 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.

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.

B. Retrieval Data Sources

1) *Manual Catalog Annotation*: 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 Fig. 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 program such as identification codes, copyright owners, available formats, and the program title.

2) *Multimedia Content Analysis*: 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 identified in Section II. For our transcripts, we use Dutch-language automatic speech recognition transcripts from the SHoUT [17] and Limsi [9] systems, as well as automatic machine translation of the Dutch SHoUT transcripts into English from the QMUL system [4]. Our low-level features are produced by the MediaMill [43] system. Finally, we utilize a lexicon of 54 concept detectors, once again from the MediaMill system, which includes detectors for concepts such as *dog*, *mountain*, *hand*, and *flower* [43]. In contrast

to the manual catalog annotations, all of the multimedia analysis sources contain noise, but are abundant and available at the shot level.

C. Query Definitions

As illustrated in Fig. 1, four query sets and their associated relevance judgments are being considered at the shot level and at the program level. This allows for evaluation of the video retrieval tasks from different perspectives: 1) current practice in the archive; 2) current content-retrieval benchmarks; and 3) a future practice in the archive, incorporating content-based search. Also, we include a set of simulated queries which allow for large scale evaluation of retrieval.

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 [23], [35]. 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 [16].

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 third and fourth examples in Fig. 3, we mark all shots within that program as relevant.

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 [38], and the 2008 VideOlympics interactive retrieval showcase [42]. 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 [8]. 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) *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-based video retrieval capabilities. Today's logged archive queries and purchases are not

User Searches	Purchased Program	Purchase Keyframes (randomly selected)
- glass haanstra - haanstra	Series: Zoo Episode: People in a Wildlife Park Purchase length: 11s	
- shots f16 - saab airplane shots	Series: Zembla Episode: The Defense Orders Purchase length: 13s	
- noorderlicht on:1996-11-10	Series: Noorderlicht Episode: The Image of the Dolphin Purchase length: 25m 13s (whole program)	
- christmas	Series: Andere Tijden Episode: WWII in Amateur Movies Purchase length: 70m 4s (whole program)	

Fig. 3. 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.

necessarily well suited for evaluating content-based video retrieval. Queries regularly do not contain words describing the required video content, consisting rather of program titles or technical codes [20]. 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 analyze search sessions and use them to formulate 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 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. 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.

4) *Query Set 4: Simulated Queries*: In addition to these query sets, we generate a set of simulated queries using the simulation framework of Huurnink *et al.* [19] on the basis of query logs from the same archive. In this simulation approach, a given catalog entry is used to generate a simulated query. The associated program is then considered relevant to that query. We

use the recommended optimal simulation strategy of selecting query terms from catalog fields according to fielded priors with a TF.IDF term selection model, as this was found to result in queries and relevance judgments that most resembled those of real logged queries [19].

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 wish. However there are limitations to this approach. Namely, the simulators create relevance judgments at the level of an entire program, and are therefore not suitable for evaluating shot-level retrieval. Furthermore, there is only one relevant program per query. 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 2190 simulated purchase-query pairs.

D. Video Retrieval Tasks

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

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.

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-based video retrieval, we employ an approach from the domain of *passage retrieval* [36] as described in Section II. We evaluated a number of approaches from the passage retrieval literature, and found the decay-based method [55] to work well in aggregating shot-level results for program retrieval.

TABLE I
STATISTICS OF THE FOUR 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
Simulated	2,190	<i>n/a</i>	2,190

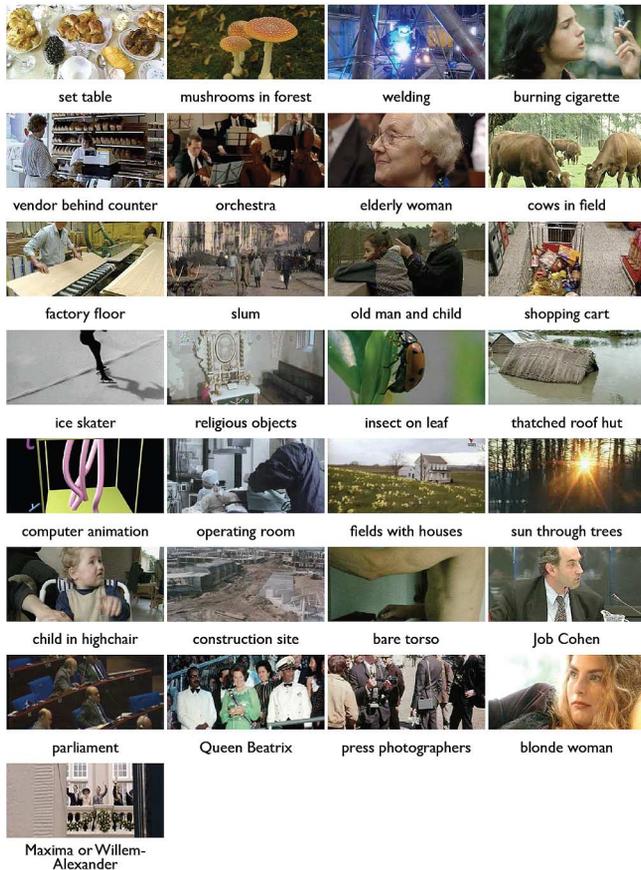


Fig. 4. Visual overview of the *future query set*, which we derived by reviewing the logged behavior of users in the audiovisual archive.

IV. EXPERIMENTAL SETUP

Now that we have outlined our evaluation methodology, we move on to describe the experimental setup. We summarize the statistics of our four query sets and their associated relevance judgments in Table I. A visual overview of the future query set, which we created by analyzing visual information needs in the archive search logs, is given in Fig. 4.

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 program was first broadcast in 1927, the most recent in 2004. The broadcasts are in the Dutch language with incidental occurrences of other languages, mostly in interviews. When this occurs, the speech transcription fails. Better transcription may be obtained

with a multiple language system. The video collection has been pre-segmented into 35 766 shots [33].

A. 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 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 RQ1 and RQ2.

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

In this experiment, we address the current retrieval practice in the audiovisual archive. We retrieve videos as complete productions using four 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*

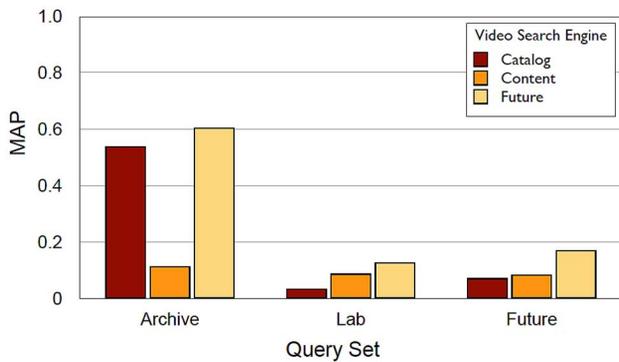
We examine the potential contribution of three different types of content-based search, namely transcript-based search, feature-based search, and detector-based search. This experiment aims at giving more detailed information for answering RQ1, RQ2, and RQ3. We perform this experiment on the queries 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 1000 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.

B. 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 [29] the query terms. Retrieval is done using the language modeling paradigm [34]. To compensate for data sparseness and zero probability issues, we interpolate document and collection statistics using Jelinek-Mercer smoothing [60]. 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.

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 [43]. Since both the detector- and feature-based retrieval methods rely on multimedia query examples as input,



Query set	Video search engine		
	Catalog	Content	Future
Archive	0.539	0.113 ∇	0.605 \blacktriangle
Lab	0.034	0.087 \blacktriangle	0.127 \blacktriangle
Future	0.071	0.084 \circ	0.170 \blacktriangle

Fig. 5. Results for Experiment 1: shot retrieval in the audiovisual archive, showing MAP scores for three query sets using three video search engines. \blacktriangle , ∇ , and \circ , respectively, indicate that a score is significantly better, worse, or statistically indistinguishable from the score using the catalog-based video search engine. A graphical representation is included to highlight patterns in retrieval behavior.

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.

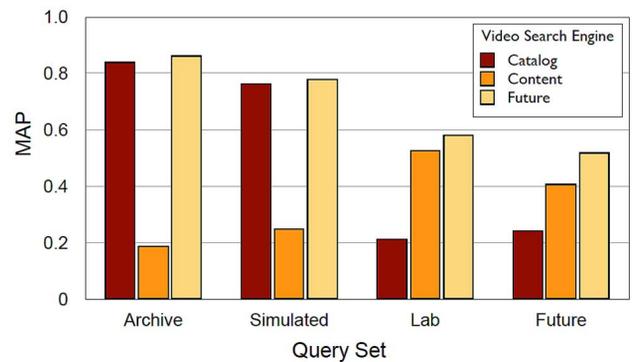
Result fusion. All three video search engines produce multiple search result that must be combined for a final retrieval outcome. Since we are concerned with evaluating the *potential* of video retrieval in the archive, we simply take for each query the combination that optimizes retrieval performance. We perform fusion using the settings recommended by Wilkins [54], i.e., we truncate each retrieval result to contain no more than 5000 items, we normalize the scores using Borda rank-based normalization, and we fuse all results using the weighted CombSUM method.

V. RESULTS

A. Experiment 1: 3×3 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 Fig. 5.

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,



Query set	Video search engine		
	Catalog	Content	Future
Archive	0.840	0.188 ∇	0.863 \circ
Simulated	0.763	0.250 ∇	0.780 \blacktriangle
Lab	0.213	0.528 \blacktriangle	0.582 \blacktriangle
Future	0.243	0.408 \blacktriangle	0.519 \blacktriangle

Fig. 6. Results for Experiment 2: program retrieval in the audiovisual archive, showing MAP scores for four query sets using three video search engines. Note the inclusion of Simulated queries, each of which is associated with exactly one relevant program.

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-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 **RQ2**, *What can content-based video 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-based video retrieval into today's practice. After combination, tomorrow's Future queries gain even more, with performance more than doubling.

B. Experiment 2: 3×4 Program Retrieval

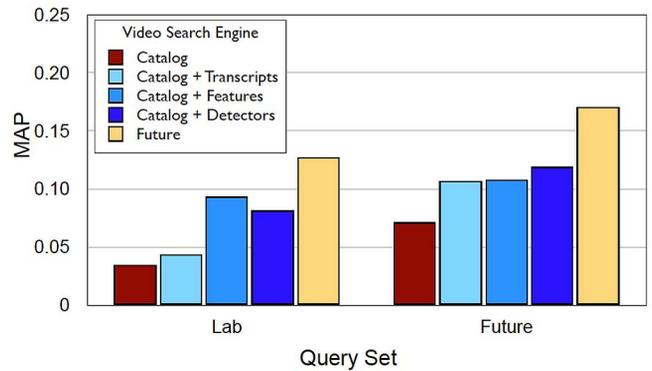
The results of Experiment 2, i.e., program retrieval with three video search engines using four query sets, are given in Fig. 6.

As was the case for shot retrieval, the Archive queries are much less responsive to the Content video search engine than the Lab and Future queries. The Archive queries gain a high absolute MAP score of 0.840, respectively, with the Catalog search engine. With the Content video search engine, the score is much lower at 0.188. 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-based video retrieval. 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. The results for the Simulated queries, i.e., program retrieval for 2190 simulated purchase-query pairs, are also shown in Fig. 6. The results for the simulated queries are similar to those for program retrieval with Archive queries in that 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 [19]. 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 a MAP score of 1; in other words, for these queries, the Content engine gave the simulated purchase the highest rank.

Returning to **RQ3**, *can content-based video 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-based video retrieval for program search: this is an artefact of the methodology used to create the queries and associated relevance judgments.

C. Experiment 3: Prioritizing Content Search

The results for Experiment 3, i.e., shot retrieval with three different content-based video retrieval methods, are shown in Fig. 7. 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 give us information as to 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



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

Fig. 7. Performance in MAP for Experiment 3; shot retrieval for two (multi-media) 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.

significant difference between the three content-based video 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-based video 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.

D. Analysis of Future Queries

Here we provide a query-level analysis of the Future queries, in particular with respect to the Future search engine, in order to gain further insight into how these are affected by the different types of automatically generated content metadata.

1) *Shot Retrieval*: A breakdown of the performance of individual queries for shot retrieval with the future engine is given in Table II. The query that gains the most from the Future search engine is for *mushrooms in the forest*, which increases in AP from 0.198 to 0.625. This query especially benefits from transcripts, as the collection contains nature documentaries which discuss mushrooms in the narrative text, close to their appearance in video. The most difficult query of all is *shopping carts*, for which no search engine returns any relevant results at all. We attribute this to the fact that shopping carts are not mentioned in the transcripts, the visual diversity of scenes in which a shopping cart may appear, and the absence of a corresponding detector.

Let us further analyze the individual queries in terms of the weighting of the different types of automatically generated content metadata, starting with transcripts. Transcripts were utilized in 13 (45%) of the queries, and were the most highly weighted information source in 6 (20%) of the queries. The query where transcripts were given the highest absolute weight is *a vendor behind the counter of a store*. Here the transcripts were very

TABLE II

WEIGHT DISTRIBUTION OVER THE DIFFERENT RETRIEVAL DATA SOURCES IN THE FINAL FUTURE SEARCH ENGINE, FOR SHOT RETRIEVAL. THE HIGHEST WEIGHTS PER QUERY ARE HIGHLIGHTED IN BOLD. QUERIES ARE SORTED IN ORDER OF THEIR ABSOLUTE IMPROVEMENT Δ AP OVER THE RETRIEVAL SCORE OF THE CATALOG SEARCH ENGINE. NOTE THE DIVERSITY OF THE MOST HIGHLY WEIGHTED DATA SOURCES ACROSS THE QUERIES

Query description	AP Future	Δ AP	Optimal weight setting			
			Catalog	Transcripts	Detectors	Features
mushrooms in the forest.	0.625	0.427	0.10	0.80		0.10
a field with cows.	0.375	0.326	0.60	0.20	0.10	0.10
a meeting in the Lower House of parliament.	0.396	0.250	0.50	0.20	0.20	0.10
a burning cigarette.	0.309	0.235	0.35	0.55		0.10
Job Cohen, the mayor of Amsterdam.	0.462	0.219	0.80		0.20	
a symphony orchestra or chamber orchestra.	0.551	0.204	0.90		0.10	
a vendor behind the counter of a store	0.178	0.178		0.90	0.10	
Queen Beatrix.	0.365	0.158	0.20	0.70		0.10
a slum.	0.311	0.155	0.67	0.21		0.11
a close-up of an insect on a leaf.	0.143	0.141	0.15	0.85		
Princess Maxima and/or Prince Willem Alexander.	0.173	0.085	0.80	0.10		0.10
a large pasture with a house or farm building in the background.	0.113	0.084	0.40		0.20	0.40
a computer animation of a process.	0.083	0.079	0.10			0.90
a construction site.	0.315	0.077	0.90		0.10	
the working area in a factory.	0.141	0.058	0.90			0.10
ice skaters with (a part of) their legs visible.	0.048	0.048			1.00	
the sun or moon shining through the silhouette of a tree or branches.	0.031	0.030		0.20	0.10	0.70
a set table, with cutlery and at least one plate visible.	0.025	0.024	0.20	0.10		0.70
a group of press photographers.	0.032	0.024	0.30		0.10	0.60
a blonde woman.	0.044	0.022	0.20			0.80
religious objects such as gold crosses and statues of saints.	0.014	0.014	0.30	0.20		0.50
a small child in a child's chair.	0.023	0.011	0.60			0.40
a patient in an operating room.	0.014	0.010	0.60			0.40
an old man (gray haired or balding, with many wrinkles) with children.	0.007	0.006	0.20		0.70	0.10
a hut with a thatched roof.	0.142	0.006	0.90		0.10	
a bare torso.	0.003	0.002	0.20	0.68		0.13
a welder at work.	0.001	0.001				1.00
an elderly woman, with gray or white hair and lots of wrinkles.	0.015	0.000	1.00			
shopping carts.	0.000	0.000	<i>n/a</i>	<i>n/a</i>	<i>n/a</i>	<i>n/a</i>

useful because a video contained a skit between a shop-keeper and a client where the shop-keeper mentioned the Dutch word for shop-counter during the skit, and thus the shots from this scene were placed at the top of the results when using transcripts. Transcripts are considered useful for finding named entities, but surprisingly transcript were not useful for finding some people, for example *Job Cohen, the mayor of Amsterdam*. Here the transcript results were not used at all; rather the catalog and the detectors were used to give the final list. An analysis of the results showed that a single program contained all but one of the relevant shots in the collection. This program contained an interview with Job Cohen, and though his name was mentioned throughout the program, these mentions were not necessarily aligned with shots of Job Cohen. Detectors were more useful because these were triggered for the concepts *person* and *face*, and could be used in combination with the catalog information to bring shots of Job Cohen to the top of the result list. So even though the transcripts contained useful information in this case, the manual catalog annotations and the detector results could attain even higher accuracy.

Turning to detectors, these were utilized in 12 (41%) of the queries. They were the most highly weighted information source

for only two of the queries, which is surprising as detectors have proved helpful for answering queries in content-based video retrieval evaluations [41]. An analysis of the queries shows that there was little overlap between the desired video content and the 54 available detectors, therefore giving results of limited usefulness. For example, the query for *a blonde woman* is one where we expected detectors to be helpful, but where detectors were not utilized at all. An inspection of the results revealed that the detectors *person* and *face* were selected for this query, and these had a bias towards detecting men as opposed to detecting women. We speculate that a detector for *female person* would have led to much better results, but such a detector was not available. Instead, features were useful in identifying relevant shots, as these were especially good at identifying blonde women shown in close-up shots with pale backgrounds. Still, even with a limited set of detectors, detectors can prove useful when the correct combination is chosen. For example, the query where detectors helped the most was for *ice skaters with (a part of) their legs visible*. Here the detectors that the system selected to answer the query were *snow*, *black and white*, and *sky*. The *snow* and *sky* detectors ensured a high ranking of winter landscapes, likely to contain people skating on ice, and the black and

TABLE III
WEIGHT DISTRIBUTION OVER THE DIFFERENT RETRIEVAL DATA SOURCES IN THE FINAL FUTURE SEARCH ENGINE, FOR PROGRAM RETRIEVAL. THE HIGHEST WEIGHTS PER QUERY ARE HIGHLIGHTED IN BOLD. QUERIES ARE SORTED IN ORDER OF THEIR ABSOLUTE IMPROVEMENT Δ AP OVER THE RETRIEVAL SCORE OF THE CATALOG SEARCH ENGINE

Query description	AP Future	Δ AP	Optimal weight setting			
			Catalog	Transcripts	Detectors	Features
a blonde woman.	0.762	0.728			0.40	0.60
shopping carts.	0.508	0.508	0.29	0.38	0.33	
a slum.	0.745	0.466	0.05	0.05	0.20	0.70
a hut with a thatched roof.	0.620	0.453	0.30		0.10	0.60
a close-up of an insect on a leaf.	0.455	0.431	0.50		0.40	0.10
the sun or moon shining through the silhouette of a tree or branches.	0.435	0.378	0.20		0.50	0.30
mushrooms in the forest.	0.707	0.373	0.10		0.40	0.50
an elderly woman, with gray or white hair and lots of wrinkles.	0.511	0.361	0.20	0.30	0.20	0.30
a field with cows.	0.761	0.361	0.30	0.10	0.30	0.30
Job Cohen, the mayor of Amsterdam.	1.000	0.357	0.57	0.41		0.02
a bare torso.	0.400	0.349		0.10		0.90
a burning cigarette.	0.588	0.310	0.60	0.10	0.30	
Queen Beatrix.	0.502	0.294	0.30	0.10	0.40	0.20
the working area in a factory.	0.639	0.287	0.30	0.20		0.50
a large pasture with a house or farm building in the background.	0.538	0.280	0.30		0.20	0.50
a computer animation of a process.	0.522	0.270	0.50			0.50
religious objects such as gold crosses and statues of saints.	0.311	0.222	0.30	0.20	0.10	0.40
a set table, with cutlery and at least one plate visible.	0.289	0.214	0.50	0.10		0.40
a vendor behind the counter of a store	0.213	0.181	0.30	0.70		
a group of press photographers.	0.288	0.162	0.40	0.20	0.30	0.10
an old man (gray haired or balding, with many wrinkles) with children.	0.222	0.144		0.30	0.30	0.40
a patient in an operating room.	0.394	0.144	0.90		0.10	
ice skaters with (a part of) their legs visible.	0.157	0.139	0.05		0.95	
a construction site.	0.324	0.138	0.40	0.10	0.30	0.20
a welder at work.	0.142	0.134	0.10	0.10	0.10	0.70
a symphony orchestra or chamber orchestra.	0.833	0.083	0.69	0.09	0.07	0.16
a meeting in the Lower House of parliament.	0.740	0.074	0.20	0.80		
a small child in a child's chair.	0.431	0.073	0.60	0.30	0.10	
Princess Maxima and/or Prince Willem Alexander.	0.931	0.000	1.00			

white detector did not harm the results as they were contained in black and white documentaries.

Finally, we look at the contributions of the features. These contributed to a total of 20 (69%) of the queries, and were assigned the highest weight for 8 of the queries. The query where features are given the highest weighting² is *a computer animation of a process*. Relevant shots for this query have a very distinctive visual appearance, being specifically computer graphics. A surprising result was that features were not used for the query *a close-up of an insect on a leaf*, which also has a very distinctive appearance. An inspection of these results showed that features returned results with broad flat surfaces for this query, such as extreme closeups of faces, statues, and leaves without insects.

2) *Program Retrieval*: A query-level summary of the optimal weight setting for program retrieval with the Future search engine is given in Table III. Looking at overall results, the query that gains the most from the future search engine is *a blonde woman*, which increases from 0.034 to 0.728. Interestingly, detectors contribute a weight of 0.4 to the performance boost here,

²We exclude the query *a welder at work* here as the performance increase is negligible in terms of absolute MAP.

while in the same query for shot retrieval, detectors were not used at all. This is due to the fact that, by virtue of aggregation over multiple shots, the detector results assign high rank to videos with many shots of people and faces. These are more likely to contain blonde women than videos with few shots of people and faces. The query that does not gain at all from content-based retrieval is *Princess Maxima and/or Prince Willem Alexander*, which already has an average precision of 0.931 using the catalog information alone. This is because when a video contains one of these two members of the royal family, their names have always been manually annotated, so that the Catalog engine achieves high performance.

Examining the queries in more detail, we see that the patterns for program retrieval are similar to those for shot retrieval; e.g., transcripts are highly weighted for retrieving *a vendor behind the counter of a store*, detectors for *ice skaters with (a part of) their legs visible*, and features for *a computer animation of a process*. However, we observe an additional change in that queries with a low absolute AP score for shot retrieval now regularly achieve substantial performance with the Future search engine. For example, the query *shopping carts* had no relevant results for any search engines in shot retrieval, but for

program retrieval achieves an AP of 0.508. Here the transcripts provide valuable information because the word *shopping* is recognized early in the video, even though shopping carts do not appear until 92 shots later. Detectors also provide important information because this query triggered the detectors *person* and *walking/running*, so that videos featuring people walking (for example behind a shopping cart) were ranked highly. Similarly to this query, the queries *an elderly woman, with gray or white hair and lots of wrinkles; a bare torso; a welder at work; a hut with a thatched roof; and an old man (gray haired or balding, with many wrinkles) with children* all benefit substantially (<0.1 increase in AP) from the Future search engine for program retrieval, where this was not the case for shot retrieval. From this we conclude that not only can automatically-generated metadata from the shot level be successfully used to retrieve entire videos, but also that the aggregation of the (noisy) shot-level results makes it even more effective for program retrieval than for shot retrieval.

VI. CONCLUDING RECOMMENDATIONS

In this paper we have investigated how content-based video retrieval can improve searches in the audiovisual archive. Our Future search engine combined manually created archive metadata and automatically generated content metadata. We applied the search engine to queries derived from the logged searches of media professionals. We found that for queries taken directly from a search log, content-based video retrieval was of limited use. Closer inspection confirmed that this was because search queries were being formulated in terms of the limited metadata available in the system, such as program title and broadcast date. In addition, the purchases used as relevance judgments were regularly for entire programs, so that shot-level retrieval could not be properly assessed. Therefore we asked an archive employee to act as a query creator, studying the searched from the archive's logs and reformulating them as they might be issued in an archive with content-based video retrieval capabilities. We found that for these queries, shot retrieval performance was more than doubled (140% relative improvement) by combining catalog-based video search with content retrieval search.

Furthermore, we found that content-based techniques could also help when applied to the program retrieval task, with program retrieval performance also more than doubling (170% relative improvement) when the two search types were combined. 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.

Turning to the relative merits of individual content-based video retrieval methods, we found that the methods based on transcripts, detectors, and features all resulted in approximately equal performance. There was no significant difference between any of the methods, however the methods were complementary, with the best performance being obtained when all three methods were combined. Based on these retrieval experiments alone, no individual source of information for content-based video retrieval is to be preferred over the others as all three methods gave approximately the same performance.

Our query-level analysis provided further insights: catalog information is useful for most queries, but much more effective when combined with retrieval data obtained from transcripts, detectors, and features. Automatically identifying queries for which content-based analysis is useful is a challenging area. We believe that recent work on predicting query difficulty is a promising starting point for addressing the problem [12], [15].

Our experiments have shown that content-based video retrieval aids the retrieval practice when incorporated with the manually created metadata that is already present in the archive. Hence, we recommend that audiovisual archives invest in embedding content-based video 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-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. The time has come to incorporate content-based video retrieval methods in the audiovisual archive, not as a substitute for existing methods, but in conjunction with them.

ACKNOWLEDGMENT

The authors would like to thank the Netherlands Institute for Sound and Vision for making available the transaction log data. The authors also would like to thank W. van de Heuvel and R. Aly.

REFERENCES

- [1] A. Allauzen and J.-L. Gauvain, "Open vocabulary asr for audiovisual document indexation," in *Proc. ICASSP*, 2005, pp. 1013–1016.
- [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 *Proc. ACM Multimedia*, San Francisco, CA, 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 *Proc. CBMI*, 2008, pp. 93–100.
- [4] S. Carter, C. Monz, and S. Yahyaee, "The QMUL system description for IWSLT 2008," in *Proc. IWSLT*, 2008, pp. 104–107.
- [5] L. Chaisorn, K.-W. Wan, Y.-T. Zheng, Y. Zhu, T.-S. Kok, H.-L. Tan, Z. Fu, and S. Bolling, "TRECVID 2010 known-item search (KIS) task by I2R," in *Proc. TRECVID*, 2010.
- [6] X. Chen, J. Yuan, L. Nie, Z.-J. Zha, S. Yan, and T.-S. Chua, "TRECVID 2010 known-item search by NUS," in *Proc. TRECVID*, 2010.
- [7] 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 Trans. Circuits Syst. Video Technol.*, vol. 17, no. 3, pp. 365–371, Mar. 2007.
- [8] O. de Rooij, C. G. M. Snoek, and M. Worring, "Balancing thread based navigation for targeted video search," in *Proc. CIVR*, New York, 2008, pp. 485–494, ACM.
- [9] 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 *Proc. Interspeech*, Brighton, U.K., 2009.
- [10] J. Foote, "An overview of audio information retrieval," *Multimedia Syst.*, vol. 7, no. 1, pp. 2–10, 1999.
- [11] J. Garofolo, C. Auzanne, and E. Voorhees, "The TREC spoken document retrieval track: A success story," in *Proc. 9th Text Retrieval Conf. (TREC)*, 2000, pp. 107–130.

- [12] C. Hauff, V. Murdock, and R. Baeza-Yates, "Improved query difficulty prediction for the web," in *Proc. CIKM*, 2008, pp. 439–448, ACM.
- [13] A. G. Hauptmann and W.-H. Lin, "Assessing effectiveness in video retrieval," in *Proc. CIVR*, 2005, pp. 215–225, Springer-Verlag.
- [14] A. G. Hauptmann, R. Yan, and W.-H. Lin, "How many high-level concepts will fill the semantic gap in news video retrieval?," in *Proc. CIVR*, New York, 2007, pp. 627–634, ACM.
- [15] J. He, M. Larson, and M. de Rijke, "Using coherence-based measures to predict query difficulty," in *Proc. ECIR*, 2008, pp. 689–694.
- [16] K. Hofmann, B. Huurnink, M. Bron, and M. de Rijke, "Comparing click-through data to purchase decisions for retrieval evaluation," in *Proc. SIGIR*, 2010, pp. 761–762.
- [17] M. Huijbregts, R. Ordelman, and F. de Jong, "Annotation of heterogeneous multimedia content using automatic speech recognition," in *Proc. SAMT*, Berlin, Germany, 2007, LNCS, Springer Verlag.
- [18] B. Huurnink and M. de Rijke, "Exploiting redundancy in cross-channel video retrieval," in *Proc. MIR*, 2007, pp. 177–186, ACM.
- [19] B. Huurnink, K. Hofmann, M. de Rijke, and M. Bron, "Validating query simulators: An experiment using commercial searches and purchases," in *Proc. CLEF*, Padova, Italy, 2010, Springer.
- [20] 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," *JASIST*, vol. 61, no. 6, 2010.
- [21] 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 *Proc. CIVR*, 2010, ACM.
- [22] 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*, vol. 12, pp. 42–53, 2010.
- [23] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, "Accurately interpreting clickthrough data as implicit feedback," in *Proc. SIGIR*, New York, 2005, pp. 154–161, ACM.
- [24] L. Kennedy, S. Chang, and A. Natsev, "Query-adaptive fusion for multimodal search," *Proc. IEEE*, vol. 96, no. 4, pp. 567–588, Apr. 2008.
- [25] M. Larson, E. Newman, and G. Jones, "Overview of VideoCLEF 2009: New perspectives on speech-based multimedia content enrichment," in *Proc. CLEF*, 2009, pp. 354–368.
- [26] M. Larson, M. Soleymani, P. Serdyukov, V. Murdock, and G. Jones, Eds., in *Working Notes of the MediaEval 2010 Workshop*, 2010.
- [27] M. Lux, K. Schoeffmann, M. del Fabro, M. Kogler, and M. Taschwer, "ITEC-UNIKLU known-item search submission," in *Proc. TRECVID*, 2010.
- [28] T. Mei, Z.-J. Zha, Y. Liu, M. W. G.-J. Qi, X. Tian, J. Wang, L. Yang, and X.-S. Hua, "MSRA at TRECVID 2008: High-level feature extraction and automatic search," in *Proc. TRECVID*, Gaithersburg, MD, 2008.
- [29] C. Monz and M. de Rijke, "Shallow morphological analysis in monolingual information retrieval for Dutch, German, and Italian," in *Proc. CLEF*, London, U.K., 2002, pp. 262–277, Springer-Verlag.
- [30] A. P. Natsev, M. R. Naphade, and J. Tešić, "Learning the semantics of multimedia queries and concepts from a small number of examples," in *Proc. ACM Multimedia*, 2005, pp. 598–607.
- [31] 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 *Proc. CIVR*, Heidelberg, Germany, 2006, pp. 143–152, Springer-Verlag.
- [32] J. Oomen and R. Ordelman, "Accessing audiovisual heritage: A roadmap for collaborative innovation," *IEEE Multimedia*, vol. 18, no. 4, pp. 4–10, 2011.
- [33] C. Petersohn, "Fraunhofer HHI at TRECVID 2004: Shot boundary detection system," in *Proc. TRECVID*, Gaithersburg, MD, 2004.
- [34] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proc. SIGIR*, New York, 1998, pp. 275–281, ACM.
- [35] F. Radlinski, M. Kurup, and T. Joachims, "How does clickthrough data reflect retrieval quality?," in *Proc. CIKM*, 2008, pp. 43–52.
- [36] G. Salton, J. Allan, and C. Buckley, "Approaches to passage retrieval in full text information systems," in *Proc. SIGIR*, New York, 1993, pp. 49–58, ACM.
- [37] M. Sjöberg, M. Koskela, M. Chechev, and J. Laaksonen, "PicSOM experiments in TRECVID 2010," in *Proc. TRECVID*, 2010.
- [38] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVID," in *Proc. MIR*, New York, 2006, pp. 321–330, ACM.
- [39] J. R. Smith and S.-F. Chang, "Visually searching the web for content," *IEEE MultiMedia*, vol. 4, no. 3, pp. 12–20, 1997.
- [40] C. G. M. Snoek and A. W. M. Smeulders, "Visual-concept search solved?," *IEEE Comput.*, vol. 43, no. 6, pp. 76–78, 2010.
- [41] C. G. M. Snoek and M. Worring, "Concept-based video retrieval," *Found. Trends Inf. Retrieval*, vol. 4, no. 2, pp. 215–322, 2009.
- [42] 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 MultiMedia*, vol. 15, no. 1, pp. 86–91, 2008.
- [43] C. G. M. Snoek *et al.*, "The MediaMill TRECVID 2008 semantic video search engine," in *Proc. TRECVID*, Gaithersburg, MD, 2008.
- [44] H. Tan and C. Ngo, "Fusing heterogeneous modalities for video and image re-ranking," in *Proc. ICMR*, 2011, p. 15, ACM.
- [45] O. Terris, "There was this film about . . . : The case for the shotlist," *J. Film Preserv.*, vol. 56, pp. 54–57, 1998.
- [46] X. Tian, L. Yang, J. Wang, X. Wu, and X.-S. Hua, "Bayesian visual reranking," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 639–652, 2011.
- [47] T. Tsirikia, C. Diou, A. de Vries, and A. Delopoulos, "Reliability and effectiveness of clickthrough data for automatic image annotation," *Multimedia Tools Appl.*, vol. 55, pp. 1–26, 2011.
- [48] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, pp. 1582–1596, 2010.
- [49] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271–1283, Jul. 2010.
- [50] E. M. Voorhees, "The philosophy of information retrieval evaluation," in *Proc. CLEF*, London, U.K., 2002, pp. 355–370, Springer-Verlag.
- [51] H. D. Wactlar, M. G. Christel, Y. Gong, and A. G. Hauptmann, "Lessons learned from building a terabyte digital video library," *IEEE Comput.*, vol. 32, no. 2, pp. 66–73, 1999.
- [52] 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*, vol. 10, no. 6, pp. 1085–1096, 2008.
- [53] T. Westerveld, "Using generative probabilistic models for multimedia retrieval," Ph.D. dissertation, Univ. Twente, Enschede, The Netherlands, 2004.
- [54] P. Wilkins, "An investigation into weighted data fusion for content-based multimedia information retrieval," Ph.D. dissertation, Dublin City Univ., Dublin, Ireland, 2009.
- [55] R. Wilkinson, "Effective retrieval of structured documents," in *Proc. SIGIR*, New York, 1994, pp. 311–317, Springer-Verlag.
- [56] R. Wright, "Broadcast archives: Preserving the future," in *Proc. ICHIM*, 2001, pp. 47–55.
- [57] R. Yan and A. G. Hauptmann, "A review of text and image retrieval approaches for broadcast news video," *Inf. Retrieval*, vol. 10, no. 4, pp. 445–484, 2007.
- [58] J. Yang and A. G. Hauptmann, "Exploring temporal consistency for video analysis and retrieval," in *Proc. MIR*, New York, 2006, pp. 33–42, ACM.
- [59] J. Yang, M. Chen, and A. G. Hauptmann, "Finding person X: Correlating names with visual appearances," in *Proc. CIVR*, New York, 2004, pp. 270–278, ACM.
- [60] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to information retrieval," *ACM Trans. Inf. Syst.*, vol. 22, no. 2, pp. 179–214, 2004.



Bouke Huurnink received the M.Sc. degree in multimedia information systems and the Ph.D. degree in computer science from the University of Amsterdam, Amsterdam, The Netherlands, in 2005 and 2010, respectively.

He is currently employed as a senior researcher at the Netherlands Institute for Sound and Vision, Hilversum, The Netherlands. He is also a guest researcher at the University of Amsterdam. This work was completed during his work as a postdoctoral researcher at the University of Amsterdam

(2010–2012). His research interests include log analysis and audiovisual search.



Cees G. M. Snoek (SM'01) received the M.Sc. degree in business information systems and the Ph.D. degree in computer science from the University of Amsterdam, Amsterdam, The Netherlands, in 2000 and 2005, respectively.

He is currently an Assistant Professor in the Intelligent Systems Lab at the University of Amsterdam. He was a visiting scientist at Carnegie Mellon University, Pittsburgh, PA in 2003 and at the University of California, Berkeley, CA in 2010–2011. His research interest is video and image search.

Dr. Snoek is the lead researcher of the MediaMill Semantic Video Search Engine, which is a consistent top performer in the yearly NIST TRECVID evaluations. He is a co-initiator and co-organizer of the VideOlympics, co-chair of the SPIE Multimedia Content Access conference, member of the editorial board for IEEE MultiMedia, and guest editor for the IEEE TRANSACTIONS ON MULTIMEDIA, special issue on Socio-Video Semantics. He is recipient of a young talent VENI grant from the Dutch Organization for Scientific Research (2008) and a Fulbright visiting scholar grant (2010).



Maarten de Rijke received the M.Sc. degree in mathematics, the M.Sc. degree in philosophy, and the Ph.D. degree in theoretical computer science from the University of Amsterdam, Amsterdam, The Netherlands, in 1989, 1990, and 1993, respectively.

He is a full Professor of Information Processing and Internet in the Informatics Institute at the University of Amsterdam, Amsterdam, The Netherlands. He leads the Information and Language Processing Systems group, one of the leading academic research groups in information retrieval in Europe. His cur-

rent research focus is on intelligent information access, with projects on search and discovery for social media, vertical search engines, machine learning for information retrieval, semantic search, and multilingual information. With an h-index of 42, he has published over 500 papers, and he has published or edited over a dozen books. He is the director of the University of Amsterdam's Intelligent Systems Lab (ISLA) and its Center for Creation, Content and Technology (CCCT)

Dr. de Rijke is a Pioneer personal innovational research incentives grant laureate (comparable to an advanced ERC grant), and has generated over 30MEuro in project funding. He is editor for various journals and book series, and a former coordinator of retrieval evaluation tracks at TREC, CLEF, and INEX (Blog, Web, Question answering). He is co-chair for SIGIR 2013.



Arnold W. M. Smeulders (M'79) is at the national research institute CWI in Amsterdam, The Netherlands, leading COMMIT, a nation-wide, very large public-private research program. He is also chair of IPN, the national policy committee for research in computer science. He is with the ISIS group at the University of Amsterdam for research in the theory and practice of visual search. He is co-owner of Euvision Technologies BV, a company spun off from the UvA. He was a visiting professor in Hong Kong, Tuskuba, Modena, and Cagliari.

He is associate editor of the IJCV. He was recipient of a Fulbright fellowship at Yale University. He is a fellow of the International Association of Pattern Recognition, and an honorary member of the Dutch Society for Pattern Recognition. He was general chairman of IEEE and ACM conferences on Multimedia.