UvA-DARE (Digital Academic Repository)

Search in audiovisual broadcast archives
Huurnink, B.

Citation for published version (APA):
Huurnink, B. (2010). Search in audiovisual broadcast archives

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

Download date: 07 Jan 2019
Chapter 10

Integrating Content-Based Video Retrieval into the Archive

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

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

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

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

In this chapter we answer the following research questions:

**CRQ 1** What is the performance of content-based video retrieval when answering
Chapter 10. Integrating Content-Based Video Retrieval into the Archive

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

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

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

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

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

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

The contributions of this quantitative study are four-fold:

- We evaluate the performance of content-based video retrieval for textual queries which are taken directly from the transaction logs of an audiovisual broadcast archive.

- We evaluate the performance of content-based video retrieval for multimedia queries that are developed by reformulating the information needs observed in session-level analysis of the searcher behavior recorded in the transaction logs of an audiovisual broadcast archive.

- We examine the gains that may be achieved by combining content-based video retrieval with retrieval using manually created catalog annotations maintained by the archive.

- We present a publicly available evaluation collection that includes manually created program annotations from the archive, queries based on the information needs of users from the audiovisual archive, and their associated relevance
10.1 Evaluation Methodology

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

10.1.1 Query Definitions

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

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

judgments.\(^1\)

The rest of this chapter is structured as follows. We present our evaluation methodology in Section 10.1. In Section 10.2 we outline our experimental setup. Results are presented in Section 10.3. We end this chapter with conclusions and recommendations for the archive in Section 10.4.
wish. However there are limitations to this approach. Namely, our simulators create relevance judgments at the level of an entire program, and are therefore not suitable for evaluating shot-level retrieval. In addition, the simulated queries do not necessarily reflect the needs of real users. Keeping these limitations in mind, we generate 10 simulated queries for each of the 219 programs in the Archive Footage collection, resulting in a set of 2,190 simulated purchase-query pairs. For the experiments in this chapter, we use the simulator that that was found to be best in Chapter 4.
10.1. Evaluation Methodology

10.1.2 Retrieval Data Sources

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

10.1.3 Video Retrieval Tasks

We consider two video retrieval tasks.

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

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

10.1.4 Video Search Engines

Video search engine 1: catalog-based  Our Catalog search engine indexes the catalog entries associated with the programs in the collection. The (Dutch language) free text, tags, and technical metadata are each indexed and retrieved separately. We normalize, stem, and decompound [109] the text. Retrieval is done using the language modeling paradigm [122], with the Lemur toolkit implementation. To compensate for data sparseness and zero probability issues, we interpolate document and collection statistics using Jelinek-Mercer smoothing [197]. In addition, as the collection of 219 catalog entries (“programs”) provides a relatively small sample from which to estimate collection statistics, we augment these with collection statistics from a sample of 50,000 catalog entries randomly selected from the archive. The
Chapter 10. Integrating Content-Based Video Retrieval into the Archive

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

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

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

**Result fusion**

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

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

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

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

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

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

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

**Performance measure and significance tests** As detailed in Chapter 6, for all four experiments, we evaluate the top 1,000 ranked shot- or program-level results using the standard mean average precision (MAP) measure. In addition, we perform Wilcoxon Signed Rank tests at the 0.01 level for significance tests.
Chapter 10. Integrating Content-Based Video Retrieval into the Archive

Experiment 1: 3x3 Shot Retrieval

<table>
<thead>
<tr>
<th>Query set</th>
<th>Video search engine</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Catalog</td>
<td>Content</td>
<td>Future</td>
</tr>
<tr>
<td>Archive</td>
<td>0.539</td>
<td>0.113*</td>
<td>0.605*</td>
</tr>
<tr>
<td>Lab</td>
<td>0.034</td>
<td>0.087*</td>
<td>0.127*</td>
</tr>
<tr>
<td>Future</td>
<td>0.071</td>
<td>0.084*</td>
<td>0.170*</td>
</tr>
</tbody>
</table>

Experiment 2: 3x3 Program Retrieval

<table>
<thead>
<tr>
<th>Video search engine</th>
<th>Catalog</th>
<th>Content</th>
<th>Future</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.840</td>
<td>0.188*</td>
<td>0.863*</td>
<td></td>
</tr>
<tr>
<td>0.213</td>
<td>0.528*</td>
<td>0.582*</td>
<td></td>
</tr>
<tr>
<td>0.243</td>
<td>0.408*</td>
<td>0.519*</td>
<td></td>
</tr>
</tbody>
</table>

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

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

10.3 Results

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

10.3.1 Experiment 1: 3x3 Shot Retrieval

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

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

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

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

In answer to CRQ 1, *What is the performance of content-based video retrieval when answering today’s queries in the archive, and queries as they might be formulated in the archive of the future?*, content-based video retrieval alone is not enough to satisfy the needs of today’s archive users. However, if future users state their information needs in content-based video retrieval terms (as is the case for the Future queries) then both search engines perform equally well. We gain the most when combining content-based video retrieval with retrieval using the catalog entries — which brings us to CRQ 2, *What can content-based video retrieval add to search performance when combined with current archive search capabilities?* Today’s Archive queries, though less sensitive to content-based methods than other query sets, gain a significant performance increase by embedding content-based video retrieval into today’s practice. After combination, tomorrow’s Future queries gain even more, with performance more than doubling.
Chapter 10. Integrating Content-Based Video Retrieval into the Archive

10.3.2 Experiment 2: 3x3 Program Retrieval

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

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

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

10.3.3 Experiment 3: Program Retrieval with Simulated Queries

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

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

10.3.4 Experiment 4: Prioritizing Content Search

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

10.4 Conclusions and Recommendations

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

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

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

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