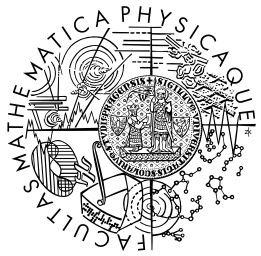


FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

DOCTORAL THESIS

Petra Galuščáková

Information retrieval and navigation in audio-visual archives

Institute of Formal and Applied Linguistics

Supervisor of the doctoral thesis: doc. RNDr. Pavel Pecina Ph.D.
Study programme: Computer Science
Study branch: Mathematical Linguistics

Prague 2017

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In . . . . . .

Title: Information retrieval and navigation in audio-visual archives

Author: Petra Galuščáková

Department: Institute of Formal and Applied Linguistics

Supervisor: doc. RNDr. Pavel Pecina Ph.D., Institute of Formal and Applied Linguistics

Abstract: The thesis probes issues associated with interactive audio and video retrieval of relevant segments. Text-based methods for search in audio-visual archives using automatic transcripts, subtitles and metadata are first described. Search quality is analyzed with respect to video segmentation methods. Navigation using multimodal hyperlinks between video segments is then examined as well as methods for automatic detection of the most informative anchoring segments suitable for subsequent hyperlinking application. The described text-based search, hyperlinking and anchoring methods are finally presented in working form through their incorporation in an online graphical user interface.

Keywords: Multimedia retrieval, audio-visual archives, video search, hyperlinking

Contents

Acknowledgement

I would like to thank all the people who contributed to the work described in this thesis. First I would like to thank my advisor Pavel Pecina for his support, advice and useful comments. Additionally, I would like to thank Shadi Saleh for his help with design and programming of the graphical user interface. I would also like to thank to Jakub Lokoč and Martin Kruliš from the Department of Software Engineering, Jan Čech from the Czech Technical University and Michal Batko and David Novák from the Masaryk University for useful discussions and help with data processing and preparation. I am very grateful to Robert Valenta for all his comments and helpful feedback. I would also like to acknowledge MediaEval organization team, especially organizers of the Search and Hyperlinking Task and Similar Segments in Social Speech Task. Foremost I would like to thank to my family for their patience and all support.

This work was supported by the Charles University Grant Agency (GA UK n. 920913), the Czech Science Foundation (grant n. P103/12/G084), the Ministry of Culture of the Czech Republic (project AMalach, program NAKI, grant n. DF12P01OVV022), Ministry of Education of the Czech Republic (project LINDAT-Clarin, grant n. LM2010013) and by SVV projects n. 260 104, n. 260 224 and n. 260 333.

1

Introduction

According to recent studies, 80% of all internet traffic in 2019 will be in video format (Wangphanitkun, 2015). At this level it would take a single person 5 million years to view all the video content crossing the internet network during one month (Cisco, 2015). This digital traffic is mainly generated by streaming services such as Netflix[1], Hulu Plus[2] and Amazon Prime[3] followed by video sharing sites such as YouTube[4], Vimeo[5] and Vine[6] (Spangler, 2015). Each minute, there are 77,160 hours of videos streamed by Netflix, more than a million videos played by Vine users (Domo, 2015) and 300 hours of videos uploaded to YouTube (Domo, 2015). This amount rose from 6 hours in 2007, to 24-35 hours in 2010 and it was expected to reach 500 hours in 2016 (Xiang, 2015). The average time spent by US adult viewers watching digital video content rose from 39 minutes per day in 2011 to almost 2 hours in 2015 (Walgrove, 2015).

The steep rise in internet video demands novel approaches to navigation throughout available video archives and necessitates robust methods to make the enormous amount of information stored in the archives readily and easily accessible to users. Services such as Netflix and YouTube even offer video recommendations, which suggest video recordings of interest to the archive users, typically based on the user's behaviour (Xiang, 2011). YouTube also provides a search engine, which is the second largest search engine in the world (Wasserman, 2014). These existing methods, however, are very functionally specific, usually

[1] `https://www.netflix.com`

[2] `https://www.hulu.com`

[3] `https://www.primevideo.com`

[4] `https://www.youtube.com`

[5] `https://vimeo.com`

[6] `https://vine.co`

relying on user-generated content, and thus are not suitable the many varied types of archives and user needs.

For example, the area of e-learning is growing rapidly with a large number of courses, lessons, and conference presentations being published online (e.g. Coursera[7], EdX[8]). Television broadcasting companies provide access to their archives and streams (e.g. BBC, CBS, ABC), which among other things also contain many news items. Moreover, there are also large numbers of private archives. Companies increasingly record their meetings and teleconferences. Users of these types of archives will more often be searching for particular information or doing an exploratory search, for which the traditional video search approaches are usually unsuitable.

This thesis documents my investigations of various approaches to high-precision navigation of audio-visual archives. Specifically, I have detailed to explore methods for searching for a particular type of information, for example, news regarding a specific event and information or details of a particular person or object. Even though the entertainment value of the video may be important, my main interest lies in identifying methods for information access. The focus is on reducing the time and effort users need to find the information of their particular interest in video archives. The large quantities of data stored in video archives today predict and identify this problem as being crucial.

Information Retrieval

The task of Information Retrieval (IR) involves searching for relevant information in the archives and extracting particular documents corresponding to a given query from these data. Manning et al. (2008a, p. 1) define IR as finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large archives (usually stored on computers).

IR is a fundamental task which enables users to find relevant documents not only in the archives but also on the internet. The information need of the user is expressed by the query, which is usually submitted as typed text. Documents relevant to the query are then retrieved and may be ranked according to their degree of relevance. IR methods usually work with structured data archives which contain texts. However, it is expected that about 80% of newly published data is not structured (Andriole, 2015). The unstructured data primarily include multimedia images, audio and video recordings.

[7] `https://www.coursera.org`

[8] `https://www.edx.org`

Accurate retrieval from multimedia archives requires a specialized form of IR, called *Multimedia Retrieval*. Multimedia Retrieval efficiently deals with the unique complexities of extracting relevant data items from multimedia archives.

Multimedia Retrieval

Multimedia Retrieval can be understood in a broader sense as a set of methods for understanding information stored in the various media in a manner comparable with human understanding (Eidenberger, 2012). Semantic content of the documents needs to be at first mined using some type of automatic processing, for example Automatic Speech Recognition (ASR), acoustic processing, image processing, face recognition, signal processing, or video content analysis. The type of processing depends on the media and information types requested by the query.

The primary focus of this work is processing videos. Videos contain different modalities, each of which require unique processing: static images can be treated by concept detection or face recognition, video dynamics can be processed by motion detection or object tracking, and the audio track, which may contain music, speech or other sounds and noise, may be processed by speech or music recognition systems. Using ASR, the textual modality of the video can also be acquired. But compared to texts, which typically have a structure defined by the author (chapters, sections, paragraphs, …) no such a structure is given in recordings. The structure of recordings is linear and compared to texts, the recordings are harder to skim, which is a problem, especially in the case of long recordings. To some extent, the structure of the recordings could be derived from audio and visual features (e.g. shots, length of silence), but such a structure still differs from a text structure.

Main objectives of the thesis

The main objective of this thesis is to improve the way how users search and navigate in large video archives. Better search methods would not only allow users to work with video archives in more productive ways, but also to find information which are not available with currently used methods. Better search methods may also help archive owners and producers to better understand their data and researchers to work with data in novel ways.

Search quality can be improved by enhancing better methods and by optimizing parameters of these methods. Such optimization should lead to more precise results and results which better correspond with user search intentions. Therefore, an important part of this thesis is devoted to experiments. I experimented

with different modalities, methods, features and parameter settings. This kind of research is extremely important in multimedia where different modalities are involved and different processing methods can be used. Number of possible ways how to improve retrieval is thus huge. Search experience can also be improved by using different search paradigm. Therefore, I also experiment with different navigation frameworks. Apart from more traditionally used text-based search, I also experiment with navigation using hyperlinks and I test highlighting informative and interesting video segments.

I show several specific ways how video retrieval can be improved and some promising areas which should be further explored. Even though some of the tested approaches did not improve the quality of retrieval, I believe that presenting these experiments can be still helpful for multimedia researches as this kind of research is unique and demanding to conduct.

Main contributions of the thesis

In this work, I document my studies of search and navigation of audio-visual archives in several ways. First, I discuss search using a traditional textual query (Chapter 2). I focus on the importance of segmentation of the recordings to be able to retrieve particular relevant passages of the recording (Chapter 3). Specifically, I explore text-based retrieval in English semi-professional television dataset.

Next, I discuss navigation in recordings using links between related segments, including situations where this type of navigation is again applied to different datasets (Chapter 4). I also describe methods for automatic selection of anchoring video segments which are supposed to be informative and interesting for the users of the archives, and thus can be used to further simplify navigation in video (Chapter 5). Last, I describe user interfaces which can be used for retrieval and navigation tasks in recordings (Chapter 6).

Thus, the main contributions of this work are as follows:

1. Providing an overview of the methods used for searching for specified information in video archives.

2. Describing a state-of-the-art, text-based search engine for searching for relevant segments of videos from video archives.

3. Describing a state-of-the-art system for linking segments of videos in archives based on their video content.

4. Comparing the relative performance of several segmentation strategies appropriate for use in searching for a particular relevant video segment.

5. Comparing the relative performance of several navigation methods on various datasets.

6. Proposing a user interface for search and navigation of video archives.

Approaches for text-based search, linking segments of videos and for automatic selection of anchoring segments described in this thesis were submitted to several shared tasks and officially evaluated by the task organizers. It was thus possible to compare our system with systems created by other teams participating in the task and working on the same research problems. Our system for text-based search ranked first in the MediaEval Benchmark Search and Hyperlinking (SH) Task 2012 (Eskevich et al., 2012a) and in the SH Task 2014 (Eskevich et al., 2014a). Our system for hyperlinking related segments ranked first in the MediaEval SH Task 2014 (Eskevich et al., 2014a).

Published work

Most of the original content described in this work has been already published in various research papers. Many published experiments were performed and evaluated in formally defined tasks organized at the MediaEval Benchmark: SH Task 2012 (Galuščáková and Pecina, 2012), SH Task 2013 (Galuščáková and Pecina, 2013a), Similar Segments in Social Speech (SSSS) Task 2013 (Galuščáková and Pecina, 2013b), SH Task 2014 (Galuščáková and Pecina, 2014b; Galuščáková, Kruliš, Lokoč, and Pecina, 2014), and Search and Anchoring in Video Archives (SAVA) Task 2015 (Galuščáková and Pecina, 2015). The experiments were described in the corresponding working note papers. The hyperlinking experiments were also examined at the TRECVID Benchmark and described in the report (Galuščáková, Batko, Kruliš, Lokoč, Novák, and Pecina, 2015).

Some of the segmentation approaches were explored in the collaborative paper (Eskevich, Jones, Aly, Ordelman, Chen, Nadeem, Guinaudeau, Gravier, Sébillot, De Nies, Debevere, Van de Walle, Galuščáková, Pecina, and Larson, 2013) published at the ACM International Conference in Multimedia Retrieval 2013. Segmentation was further explored in the paper (Galuščáková and Pecina, 2014a) published and presented at the ACM International Conference in Multimedia Retrieval 2014. The experiments with speech retrieval and acoustic processing were published (Galuščáková and Pecina, 2015) and presented at the Workshop on Speech, Language and Audio in Multimedia organized at the ACM Multimedia Conference 2015. Visual descriptors were more closely described in the

paper (Galuščáková, Batko, Čech, Matas, Novák, and Pecina, 2017) published the ACM International Conference in Multimedia Retrieval 2017. User interface for retrieval was published online (Galuščáková et al., 2016) and described in the demo paper (Galuščáková, Saleh, and Pecina, 2016) published and presented at the European Conference on Information Retrieval 2016.

2

Text-based search in multimedia

As previously mentioned, an optimal approach for a given multimedia retrieval system must be based on the modality and form of the stored information and the type of query. Defining the input format of the query becomes crucial and influences the user's comfort, satisfaction, time needed to find required information and even the possibility of successfully finding it. Systems primarily used for image retrieval may allow users to find pictures similar to a submitted query image (Flickner et al., 1995). Some systems allow a user to sketch the input query image or define colors used in different parts of the image (Blažek et al., 2015; Kuboň et al., 2016). Other video retrieval systems even enable specifying the type of using flow fields (Rossetto et al., 2016). In the case of audio retrieval, it is possible to submit a recording of a part of a song (Wang, 2003), or even input singing or humming (Ghias et al., 1995).

But for most users, the preferred manner of searching for specific information is to type a query (Levene, 2010, p. 195). This type of navigation is normally used by IR systems to search in text archives, but it is also frequently used in multimedia retrieval systems (Rasiwasia et al., 2010). However, because of several modalities present in the video, the ambiguity of the query can be huge. A typed query may not only denote an event or object mentioned in the dialog, but also match a person, object or place occurring in the video, or may even match visual text appearing in the video or music playing in the background.

Text-based search

In this thesis, *text-based search* denotes the above described archival navigation, in which the input query is typed by the user. This type of navigation can be applied either to textual, speech, audio, or video modalities. This thesis describes content-based methods, which not only rely on available metadata describing the documents, but which are able to analyse the content of the doc-

uments in the archive. The core system uses ASR to analyse the audio track. Sometimes manual or partially manual transcripts or subtitles are available and can be used instead. IR methods are then run on the acquired texts and the best matching documents are retrieved. The quality of this retrieval will depend on the quality of the ASR system and will be influenced by existing ASR problems (restricted vocabulary, recording quality, background noise, differences in pronunciation, accents, missing punctuation, etc.). Finally, IR can be enhanced by incorporating additional modalities.

Keyword Spotting and Spoken Term Detection

Some of the text-based search systems use Keyword Spotting (Moyal et al., 2013) and Spoken Term Detection (National Institute of Standards and Technology, 2006) methods instead of IR methods.

In contrast to IR, documents must contain the exact word or phrase (sometimes different word forms are also permitted) to be marked as relevant using these methods. No other information about the degree of relevancy is available. In IR, the query corresponds more closely to the topic and documents are retrieved according to their relevance to this topic. Queries thus may be drafted more freely. For example, consider a user searching for the word *bulletproof*. In a typical case of Keyword Spotting and Spoken Term Detection, all documents containing this word will be retrieved. Therefore, the sentence, "*Rover finds bulletproof evidence of water on early Mars.*" will be retrieved. In the case of IR, the input query corresponds to the query topic and the retrieval methods are intended to retrieve articles more closely corresponding to this topic on higher ranks of the list of retrieved results. Therefore, the article containing the sentence, "*A bulletproof vest is an item of personal armor that helps absorb the impact from firearm-fired projectiles.*" should be ideally retrieved on higher ranks.

Keyword Spotting and Spoken Term Detection are more robust in dealing with problems related to ASR. Both approaches can tolerate out-of-vocabulary words by processing subword units within these words. In Keyword Spotting, the list of keywords is known before speech recognition is initiated. Spoken Term Detection is an open vocabulary problem. Another problem related to these tasks is Spoken Query Retrieval (Barnett et al., 1997). In this case, the query is not typed by the user, but it is input as recited speech and recorded by the system.

Chapter overview

This chapter deals with text-based search. First, multimedia archives will be described, including datasets in which retrieval methods have already been

applied and web archives in general on which multimedia retrieval methods can possibly be applied in the future. Evaluation methods used for an assessment of the text-based search will then be described. Finally, overall system and retrieval model tuning will be described with experiments using text-based search.

2.1 Multimedia archives and datasets

The number and sizes of popular audio and video archives rose rapidly during the past few years, mainly due to better accessibility of recording devices, cheaper and faster internet connections and more widely accessible broadband. The amount of data uploaded to YouTube per minute rose from 48 hours in 2013 to 300 hours in 2015 (Domo, 2013, 2015). Most online videos accounts licensed original content provided by services like Netflix, Amazon Prime or Hulu Plus, with Netflix accounting for 36.5% of all downstream internet bandwidth during peak periods in North America (Spangler, 2015). Huge amounts of content is also provided by user content sharing services like YouTube, which accounted for 15.6% of downstream internet traffic in 2015. Large numbers of videos are also contained in various television archives (e.g. BBC Archive[1], Vanderbilt Television News Archive[2]). Videos are now more often provided for various lessons (e.g. Coursera, Khan Academy[3], Lynda[4]), lectures (e.g. VideoLectures.NET[5], Academic Earth[6]), and conferences (e.g. TED Talks[7]).

Also in existence are datasets created specifically for preserving and providing access to historical information and events. These archives include the Internet Archive[8] which is a non-profit digital library which provides access to not only large numbers of books, articles and mined webpages, but also movies, video recordings and audio recordings. Specifically, the video archive contains several different types of content, including TV news, various television programmes, comedies and silent films. The audio archives include content such as the LibriVox Free Audiobook archive, radio programmes, and podcasts. Other archives

[1] http://www.bbc.co.uk/archive

[2] https://tvnews.vanderbilt.edu

[3] https://www.khanacademy.org

[4] https://www.lynda.com

[5] http://videolectures.net

[6] http://academicearth.org

[7] https://www.ted.com

[8] https://archive.org

created to preserve and provide information include, for example, Visual History Archive[9], Critical Past[10] or UCLA Film & Television Archive[11].

Datasets for multimedia retrieval

The professional historical archives cited above are typically not available for direct download and further processing and annotation would be needed before being able to use them for research or experimental purposes. Archives available for direct downloads, which may be used for IR experiments, are called datasets for the purposes of this thesis. These datasets include AMI Meeting dataset (Carletta et al., 2006), The TDT dataset (Cieri et al., 2002), MED Summaries dataset (Potapov et al., 2014) and EVent VidEo dataset (Revaud et al., 2013).

The AMI Meeting dataset consists of 100 hours of annotated meeting recordings and has previously been used in meeting retrieval experiments (Eskevich and Jones, 2014). The TDT dataset was proposed for topic detection and tracking experiments and includes audio broadcast news and story segmentation. The MED Summaries dataset contains 160 one to five minute long videos with a manually input summary of each video. The EVent VidEo dataset, containing almost 3000 YouTube videos of 13 different events, was proposed for experiments with automatic event detection.

Experiments presented in this work were conducted on three different datasets: Blip.tv dataset (Eskevich et al., 2012c), a dataset of recorded interviews provided in the Similar Segment in Social Speech Task at the MediaEval Benchmark (Ward and Werner, 2013) and a dataset of BBC TV broadcasts (Eskevich et al., 2013b). An overview of the datasets is shown in Table 2.1.

<table>
<tr><th>Task</th><th>Train Set</th><th>Test Set</th></tr>
<tr><td>MediaEval 2012 SH</td><td>Blip.tv train set</td><td>Blip.tv test set</td></tr>
<tr><td>MediaEval 2013 SH</td><td colspan="2">BBC 2013 SH set</td></tr>
<tr><td>MediaEval 2014 SH</td><td rowspan="2">BBC 2013 SH set</td><td rowspan="2">BBC 2014 SH test set</td></tr>
<tr><td>TRECVID 2015 Video Hyperlinking</td></tr>
<tr><td>MediaEval 2015 SAVA</td><td>BBC 2015 SAVA train set</td><td>BBC 2015 SAVA test set</td></tr>
<tr><td>MediaEval 2013 SSSS</td><td>SSSS train set</td><td>SSSS test set</td></tr>
</table>

Table 2.1: An overview of the datasets used in this thesis.

[9] https://sfi.usc.edu/vha

[10] http://www.criticalpast.com

[11] https://www.cinema.ucla.edu

Blip.tv dataset

The Blip.tv dataset (Schmiedeke et al., 2013) was provided in the SH Task at the MediaEval 2012 Benchmark (Eskevich et al., 2012b). The provided videos containing semi-professional video content were collected from the Blip.tv website[12] and were published under the Creative Common license. These videos vary significantly in format (e.g. local television news, interviews, culinary shows, personal blogs), length, quality and even the language. The predominant language is English, but the dataset also contains videos in other languages, for example in French, Spanish, German and Dutch.

	Train Set	Test Set
Number of documents	5288	9550
Hours of video	1143.2	2144.6
LIMSI sentences	369 k	457 k
LIUM speech segments	350 k	705 k

Table 2.2: Size of the Blip.tv dataset.

The dataset is divided into train and test sets (Table 2.2). Each recording is published with two transcripts created by the LIMSI/Vocapia (Lamel and Gauvain, 2008) system and the LIUM system (Rousseau et al., 2011), metadata, shot boundaries (Kelm et al., 2009), face clustering, and visual concepts. The LIUM transcripts consist of one-best hypothesis, word-lattices, and confusion networks and the LIMSI transcripts include word variations with their confidence scores. Even though the language of the video varies, the LIUM system only transcribes into English. LIMSI transcripts could be in several languages. The LIMSI system first detects the language of the recording before processing it, and then transcribes into detected language. In the LIMSI transcripts, segmentation into sentences is available and the transcripts are divided into speech segments; each speech segment corresponds to a continuous utterance of one speaker. The subtitles also contain additional information and may thus also include unuttered words, such as "*SCHOOL BELL RINGS*", "*SNAP*", or song lyrics: "*# BOTH: It's wo-o-o-o-onderland... #*".

BBC datasets

The BBC dataset was provided in the 2013-2014 SH Task (Eskevich et al., 2013b, 2014b), in the 2015 SAVA Task (Eskevich et al., 2015) and in the 2015 Video Hyperlinking Task (Over et al., 2015).

[12] Blip.tv was acquired by Maker Studios in 2013 and shut down in 2015.

A single set was published in the 2013 SH Task and used for both training and testing purposes. This set is further referred to as *BBC 2013 SH set*. Even though the same set was used for training and testing, a new set of queries was introduced for testing. BBC 2013 SH set was also used for training purposes in the 2014 SH Task. But for the testing purposes in the 2014 SH Task, a new test set and a new set of queries were published. This newly published test set is further referred to as *BBC 2014 SH test set* and together with BBC 2013 SH they are further referred to as *BBC 2014 SH set*. The BBC 2014 SH set was also used in 2015 Video Hyperlinking Task, except a new set of queries which was used for testing. The anchoring sub-tasks of the 2015 SAVA Task used a sub-set of the BBC 2014 SH set. 42 videos out of 5380 were selected and used for training and they are further referred to as *BBC 2015 SAVA train set* and 33 videos were selected for testing they are further referred to as *BBC 2015 SAVA test set*.

The BBC 2013 SH set consists of BBC TV programmes broadcast between 01.04.2008 and 11.05.2008, and the BBC 2014 SH test set consists of programmes broadcast between 12.05.2008 and 31.07.2008. Therefore, the variability of the dataset is high. Among other things, it contains news (e.g. *BBC Breakfast*, *BBC News at Ten*), documentaries (e.g. *The Life of Mammals*, *Wild China*), serials (e.g. *EastEnders*, *Two Pints of Lager and a Packet of Crisps*), entertainment programmes (e.g. *Top Gear*, *Hard Sell*), quiz shows (e.g. *Eggheads*, *The Weakest Link*), cookery shows (e.g. *Saturday Kitchen*, *Ready Steady Cook*), sport events (e.g. football matches, horse races), and children's programmes (e.g. *Big Barn Farm*, *Nina and the Neurons*). The dataset also contains concerts (e.g. *Radio 1's Big Weekend*, *Glastonbury The Best Bits*) and music programmes (e.g. *Mad about Music*, *Later... with Jools Holland*) which may need special care. Thus the range of topics of the videos is unrestricted and the dataset contains many speakers, settings, accents, genres, and formality types. Additionally, the quality of the videos is very high, and the dataset contains valuable manually created metadata.

Apart from videos, the dataset also contains subtitles and three automatic transcripts provided by LIMSI (Lamel, 2012), LIUM (Rousseau et al., 2014) and NST-Sheffield (Lanchantin et al., 2013), manually entered metadata, automatically detected shots, a list of stable keyframes, prosodic features (Eyben et al., 2013), and visual concepts (Tommasi et al., 2014; Chatfield et al., 2015). Sizes of the BBC 2013 SH set and BBC 2014 SH test set, which together form the BBC 2014 SH set, are shown in Table 2.3.

	BBC 2013 SH set	BBC 2014 SH test set
Number of documents	1860	3520
Hours of video	1335	2686
LIMSI speech segments	431 k	821 k
LIUM sentences	580 k	1.1 M
NST-Sheffield segments	1.4 M	2.9 M

Table 2.3: Size of the BBC 2014 SH set.

	SSSS train set	SSSS test set
Number of documents	20	6
Hours of video	4	1
Annotated segments	1697	189
Annotated similarity sets	198	29

Table 2.4: Statistics of the SSSS data dataset.

Dataset used for the SSSS Task

The training and test data in the SSSS Task (Ward et al., 2013), further referred to as *SSSS train set* and *SSSS test set*, respectively, consist of recordings of invited interviews of two speakers, mainly computer science university students and professors participating in the survey. Participants were asked to talk about topics of interest and the conversations were recorded with the understanding that they would be further used for search experiments. Interview topics were not restricted but topics such as "*movies*" and "*university studies*" were suggested. In addition to manual transcripts created by the task organizers and automatic transcripts provided by the University of Edinburgh, the dataset contains detailed prosodic features and metadata (e.g., age, native language, gender of the speaker, and recording conditions for each document). The provided automatic transcripts are given for two tracks - one for each speaker. Therefore, we first merged these tracks into a single one, based on the time stamps of the transcribed words. Videos were published with annotations of manually indicated segments which were also manually grouped into similarity sets. Almost 1900 segments were marked in the data. Similarity sets were marked by tags such as *food*, *travel*, *planning-class-schedule*, or *family*; a total of 198 tags were assigned. The size of the dataset is stated in Table 2.4.

2.2 Evaluation

Similar to that of most textual IR systems, multimedia retrieval systems are also frequently evaluated via the Cranfield experimental setting techniques (Cleverdon and Kean, 1968) using a test dataset, manually crafted queries and manually generated ground truth data. Similar measures are also used in the evaluation process. *Precision* reports the ratio of retrieved results which are correct. If ranking of the retrieved segments is available, Precision at certain levels can also be calculated. For example *Precision at 10 (P10)* gives the ratio of correct instances from the top 10 retrieved results. *Average Precision (AP)* can then be calculated as the average of all Precision values calculated at each point when any new relevant document is retrieved (Manning et al., 2008b). This measure thus prefers systems which retrieve correct results among the top retrieved items. *Mean Average Precision (MAP)* is then calculated as the mean of AP values over the set of queries. *Reciprocal Rank (RR)* is also frequently used for the evaluation of IR systems. It is calculated as the inverse of the rank of the first correctly retrieved item. *Mean Reciprocal Rank (MRR)* (Voorhees, 1999) can then be calculated as the mean of the RR values over the set of queries. *Recall* reports the ratio of correct instances which were retrieved by the system. However, if the dataset is too large, it is often not possible to manually judge all the included results. Therefore, some adaptations of traditional evaluation measures need to be applied instead, e.g. pooling or sampling can be applied to the list of results (Over et al., 2015).

More complex system's characteristics

A full range of complex characteristics is used to acquire a more precise image of the system's performance. The most common ones are *Precision-Recall Curve*, *Receiver Operation Characteristics (ROC)*, and *Normalized Discounted Cumulative Gain (NDCG)*.

The *Precision-Recall Curve* (Manning et al., 2008b) displays the trade-off between Precision and Recall values. For most of the systems, both values should be optimized, but for some systems, Precision may be more important (e.g. in web searches since we do not want to overwhelm the user with too many distantly relevant results and we only want to show highly relevant web pages), for others higher Recall is preferred (e.g. for medical retrieval we need to retrieve all suspicious items).

ROC curves (Manning et al., 2008b) display a system's sensitivity versus its specificity. Sensitivity measures the ratio of correctly retrieved relevant documents (i.e. Recall) while the specificity measures the ratio of non-retrieved irrele-

vant documents (usually 1-specificity is used instead). ROC curves thus illustrate the "hit rate" of the system versus its "false alarm rate". Ideally, the system should have high sensitivity (high hit rate) and high specificity (low false alarm rate).

NDCG (Candan and Sapino, 2010) is based on the calculation of the *Cumulative Gain (CG)*. If each document has an assigned value representing its relevance to the query, then the *CG* can be calculated as a sum of these values over the set of retrieved documents. *Discounted Cumulative Gain (DCG)* takes into account also the ordering of the retrieved documents, with those having lower rank achieving lower emphasis. Finally, NDCG is DCG normalized to the highest DCG which can be achieved on the given set of documents.

Evaluation campaigns

Appropriate evaluation demands high quality manually created data, collected queries and manually assessed ground truth data. Assembling these kinds of data is usually very expensive and time demanding. Multimedia data is typically even harder to acquire than text. Copyright of the recordings must enable their processing by different systems and it must be able to provide recordings to human evaluators. Privacy settings must also be in the concern. The evaluation dataset and the query set should be balanced and should provide a good estimation how would the system in question work on real world data. Evaluation on multimodal data is also very time demanding and requires proper technical equipment.

Therefore, existing evaluation campaigns and benchmarks such as MediaEval[13], TREC[14], TRECVID[15], The CLEF Initiative[16], NTCIR[17], and FIRE are a great assistance. These campaigns offer sets of shared tasks in which it is possible for research teams to participate. Tasks usually provide test data, queries and a common evaluation procedure and so they offer a good opportunity to evaluate proposed methods and compare them to solutions of other participating teams. Sometimes training data and pre-processed features are also provided by the task organizers. This is again especially valuable for multimedia as processing multimodal data is time and source consuming and not always accessible. Providing evaluation framework and pre-processed features allow researches to focus on im-

[13] `http://www.multimediaeval.org`

[14] `http://trec.nist.gov`

[15] `http://trecvid.nist.gov`

[16] `http://www.clef-initiative.eu`

[17] `http://research.nii.ac.jp/ntcir/index-en.html`

portant research questions and make easier for the new researchers to start their research in the field.

Participating in shared evaluation benchmarks is crucial for this thesis as no standardized test sets, such as CoNLL-2003 for named entity recognition (Tjong Kim Sang and De Meulder, 2003) and Penn Treebank Wall Street Journal for morphological tagging (Marcus et al., 1993), are in existence for multimedia retrieval. Unified dataset, set of queries and evaluation procedure provided in the shared tasks are thus needed to be able to directly compare performance of proposed methods with other systems.

MediaEval Benchmarking

The MediaEval Benchmark (Larson et al., 2015) offers a range of multimodal tasks, such as speech search, person discovery in videos, violent scenes detection, geo-location prediction based on images, and retrieval from classical music scores. Organized tasks include the SH Task, SAVA Task and SSSS Task. These three tasks are closely related to search in multimedia and they correspond very well to the objectives of this thesis. The thesis is thus partially shaped by nature of the mentioned tasks.

In the scenario of the SH Task, a user wishes to find information relevant to a given query and then navigate through a large archive using hyperlinks to the retrieved segments. The main goal of the known-item Search sub-task is to find passages relevant to a user's interest given by a textual query in a large set of audio-visual recordings. Subsequently, the goal of the Hyperlinking sub-task is to find more passages similar to those retrieved. The SH Task thus provided us data and evaluation procedure for exploring text-based search in videos. Hyperlinking sub-task helped us to to explore alternative approaches to navigation in videos with the use of links between related video segments.

The primary objective of the SSSS Task is to find segments similar to input, or query, segments in a dataset of audio-visual recordings containing English dialogues of a university students' community. In the intended scenario, a new member (e.g., a new student) joins a community or organization (e.g., a university), which owns an archive of recorded conversations among its members. The new student would like to find information according to his or her interest in the archive to better understand the organization. The student wants to find additional video segments similar to the ones received in response to the initial query and continues to browse the archive using hyperlinks contained in each newly retrieved video. Thanks to SSSS Task, we further explored hyperlinking navigation method. Dataset with thousands of manually marked segments provided in the task was utilized to explore segmentation methods and nature of segments.

Different type of setting also enabled us to test hyperlinking methods on more diverse dataset.

The SAVA Task evolved from the SH Task. This task is specifically intended for professional and semi-professional users of TV broadcast archives. In addition to its Search sub-task, the primary intent of the Anchoring sub-task is to automatically locate segments of videos which are of interest to the archive's users. These segments are selected to stimulate users' interest in searching for additional detailed information; for example a person or place can be mentioned or an event can be represented. In this task, we further explored how can navigation in videos be simplified and how it can improve the way how do users interact with video archives. We focused on automatic detection of interesting anchoring segments which are supposed to be good source segments for hyperlinking.

MediaEval submission procedure

Benchmark tasks are organized with respect to specific research problems such as how to make automatic support for searching and hyperlinking videos or how to protect privacy in videos but still be able to keep as much useful information as possible. Research teams can choose to participate the organized tasks and work on the research problems. The task organizers then provide training data to the task participants. Training data is often provided with additional pre-processed features to enable participants to better focus on the research problem. However, provided training set is often relatively small and contains only several training examples and it thus cannot be used for training full machine learning-based system. However, training set can be well utilized for comparing different approaches on specific data and for tuning parameters of proposed methods.

The evaluation scheme differs by the task, but typically training data is provided at the beginning of the task and test data is provided about two months after the beginning of the task. Participating teams then have about a month to prepare the submissions, i.e. the results for the test queries. Participating teams can typically prepare four to five different submissions to be able to compare different settings of their systems. After the results of participating teams are submitted, task organizers evaluate the results, for example using crowdsourcing, and provide the results to the task participants.

Video Hyperlinking at TRECVID

TRECVID (Over et al., 2015) is a workshop organized by the National Institute of Standards and Technology aimed at providing test datasets and evaluation procedures for content-based video retrieval. The organized tasks include

Semantic Indexing, Multimedia Event Detection, Surveillance Event Detection, Instance Search, and Video Hyperlinking. In the Video Hyperlinking task the scenario is that users seeking more information about topics in the videos can navigate through the archive using links between related segments, jump between videos, and find detailed information about the topics of interest this way. These links should be created automatically using visual, audio and possibly metadata features.

Evaluation of retrieval of relevant segment

In the case of evaluation of retrieval of the precise relevant segment of the recording, evaluation measures commonly used for the full video retrieval need to be modified. In the Instance Search Task organized in the TRECVID, recordings are divided into film shots (i.e. uninterrupted series of frames) and each shot is marked as relevant or irrelevant. In this case, Precision- and Recall-based methods used in traditional IR can be used.

In the case of a large dataset, manual annotation of all sub-segments of recordings is often unrealistic. Also, IR systems should not only take into account whether the segment overlaps with the correct ones, but also the distance of the retrieved segment from the relevant segment. If the system retrieves a segment which begins 10 seconds before the segment marked as relevant, it is still very probable that the user can find the information of his or her interest. Therefore, it should be rewarded better than had it retrieved a segment beginning 10 minutes before the relevant segment. In such case, the probability of finding the required information is low. To deal with this problem, evaluation measures such as *Mean Reciprocal Rank Window (MRRw)*, *Mean Generalized Average Precision (mGAP)*, *Mean Average Segment Precision (MASP)*, *Binned Relevance (MAP-bin)*, *Tolerance to Irrelevance (MAP-tol)*, and *Mean Average interpolated Segment Precision (MAiSP)* may be used.

The *MRRw* (or *MRR-window*) evaluation measure is an adaptation of the *MRR* measure. Instead of the first correctly retrieved document it takes into account the segment which is correctly retrieved inside of a window with the given length lying around the relevant segment.

The *Generalized Average Precision (GAP)* (Liu and Oard, 2006) measure also employs the exact jump-in point, which represents the start of the relevant segment. In the presented experiments, it is calculated as follows:

$$\mathrm{GAP} = \frac{1}{\mathrm{rank}} \mathrm{Penalty}(\mathrm{distance}) \tag{2.1}$$

where *rank* is the rank of the fist correctly retrieved document and *Penalty* assesses the quality of the jump-in point. The Penalty value is estimated according

to the penalty function, based on the distance between the starting point of the relevant segment and the starting point of the retrieved segment. The shape of the penalty function is triangular and it depends on the given window width. In our previous work, the adapted penalty function (Galuščáková, Pecina, and Hajič, 2012) was proposed. This penalty function better corresponds with the satisfaction of users when they are searching for particular information. *mGAP* is then calculated as the average of *GAP* values over the set of the queries.

The *MASP* (Eskevich et al., 2012e) evaluation measure exploits the precision of the entire retrieved segment; i.e. both starting and ending points of the relevant segment are taken into account. *Segment Precision* is calculated as the length of the relevant retrieved segment (document correctly retrieved inside of a given window) as a fraction of the overall length of the relevant segment. *Average Segment Precision* can then be calculated as the average of *Segment Precision* values calculated for the set of relevant retrieved results and *MASP* is calculated as the average of *Average Segment Precision* values over the set of queries. *mGAP* and *MASP* are graphically illustrated in Figure 2.1.

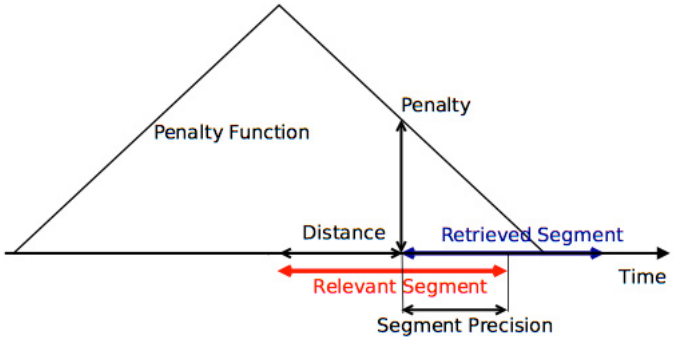


Figure 2.1: Graphical depiction of *mGAP* and *MASP* evaluation measures.

The window length is a parameter used by each described evaluation measure. In the reported experiments, windows of 60-seconds-length were used.

The *MAP-bin* and *MAP-tol* evaluation measures (Aly et al., 2013) are both adaptations of the *MAP* measure proposed for evaluation of video content retrieval to allow a segment retrieved near the relevant segment (but not necessarily overlapping it) to also be marked as relevant. In applying the *MAP-bin* measure, the recordings are split into bins of uniform length. If a retrieved segment lies in the bin with the relevant segment, it is also marked as a relevant. Only the highest ranked segment in each bin is taken into account, other returned segments lying inside of the same bin are considered to be irrelevant. If the beginning of a relevant segment lies in the tolerance window near the beginning of the retrieved segment, this retrieved segment is also marked as relevant according to

the *MAP-tol* measure. The bins in the evaluation are usually 5 minutes long, and the width of the tolerance window is usually 15 seconds. In opposition to the *MRR* measure, each relevant segment is again assigned only to the highest ranked retrieved segment in its vicinity. Using this measure thus simulates the user behaviour of not being satisfied by additional retrieved relevant segments lying close to each other (Racca and Jones, 2015). *MAP-bin* and *MAP-tol* measures are explained in Figure 2.2

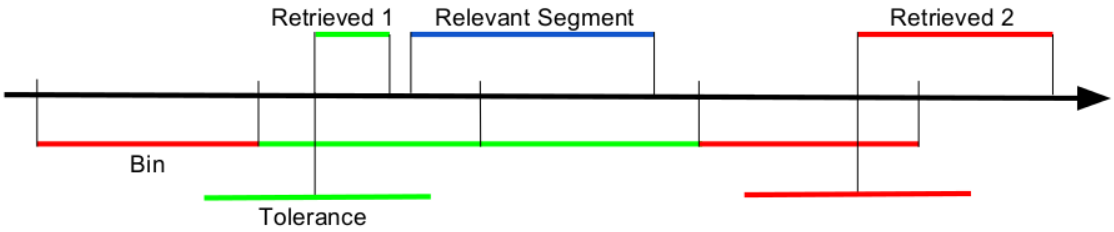


Figure 2.2: Graphical depiction of *MAP-bin* and *MAP-tol* evaluation measures.

MAiSP (Racca and Jones, 2015) reflects the time needed to find relevant content and time spent watching relevant and irrelevant content. It combines the user's effort which is measured by "the number of seconds that a user watches" and the user's satisfaction measured by the "number of seconds of new relevant content that the user can watch starting from the beginning of the segment" (Over et al., 2015). The number of relevant and irrelevant seconds listened by the user is also essential in the *Normalized Searcher Utility Ratio* (Ward et al., 2013) used in the SSSS Task.

If possible, the statistical significance of the differences between results achieved using described measures is also presented in this thesis. The significance is in all cases calculated according to the Wilcoxon signed rank test (Wilcoxon, 1945) at the 0.05 level.

Queries for text-based search

Queries used in the evaluation process should ideally correspond to real users' needs related to the dataset in question. In the SH Tasks, the textual queries were collected in surveys and by crowdsourcing.

In 2012, the queries were collected using crowdsourcing via Amazon Mechanical Turk[18]. In order to simulate the known-item search scenario, participants were first asked to find remarkable passages in the recordings and then to briefly comment on them (Larson et al., 2011). This process differs from usual query

[18] `https://www.mturk.com`

input methods in which a user first specifies the query and then judges the retrieved passages. The reverse procedure causes higher overlap of the queries and relevant passages because users tend to use the vocabulary from the recording. On the other hand, the queries tend to be more diverse. In total, 60 queries were collected; 30 were used as the train set and the remaining were was used for testing. The queries consist of "*Title*", which briefly describes the passage, "*Short Title*", which describes the passage in a more "search engine" style, information on whether the segment contains a face, the main color of the segment, and the main visual concept (e.g. Rocky Mountains, Volcanoe, Chair, Piano), if there is any. An example of a question and relevant segment is presented in Table 2.5.

Title	Profit Partner programe talks about growing business faster.
Short Title	the profit partner growing business faster mortgages
Face	Yes
Colours	Dark
Video Content	Chair, Woman
Relevant Segment Transcription	Welcome to the Profit Partner where we help you grow six figure businesses in twelve months or less. My name is Cheree Warrick and I am the Profit Partner and I am so very honoured today to the interviewing Sarah Pichardo of George Mason Mortgage.

Table 2.5: Query example used in the MediaEval 2012 SH Task

In 2013, a test set of 50 queries was created in the user study with 30 users between ages of 18 and 30. The task was again considered as known-item. Users were asked to browse the dataset, find segments of their interest and formulate text and visual queries that they would use to find these segments. The queries were not checked for spelling; some of them contained errors (e.g. "*Medieival history of why castles were first built*"), to simulate real world conditions. An example of the query used in this task appears in Table 2.6.

In 2014, the SH Task was changed from known-item to ad-hoc. 36 queries were formulated by 28 survey participants. (e.g. policeman, hair dresser, bouncer, sales manger, student) (Eskevich et al., 2014a) and the queries were designed for the home usage scenario of the system (Eskevich et al., 2014b). The users were already familiar with the dataset and were instructed to formulate queries so that there would be several relevant segments for each of them. Because of the home scenario, tablets were used in the user survey instead of computers, which were

Textual Cue	Space-Cowboys Space Pirates Pirates in Space talking music
Visual Cue	'Guisto Radio' Space Cowboys singing music Captain DJ taking calls and being foolish

Table 2.6: Query example used in the MediaEval 2013 SH Task

ID	Query	Relevant Returned	Relevant Total
2	egypt travel	19	21
10	car race	40	92
26	weight loss	15	23
31	famous cultural festival	32	55
36	gorillas wild	15	26

Table 2.7: Examples of queries used in the evaluation of the Search sub-task in the 2014 SH Task. *Relevant returned* is the number of relevant segments retrieved by the baseline system applied to available subtitles, and *Relevant total* is the number of all relevant segments of data (retrieved by any submitted system).

used in previous years. This resulted in the much shorter queries comparing with the previous years. The set of 30 queries was finally used in the Search sub-task evaluation (e.g. "*wimbledon trophy award*", "*polar bears*", "*meat recipe*"). The queries were manually edited, and only queries for which a sufficient number of relevant segments was found in the dataset were used in the Search sub-task. The queries were also spell-checked.

Several examples of the test queries used in the 2014 Search sub-task are displayed in Table 2.7, including the number of relevant segments and relevant segments retrieved by the presented system. The presentation of the retrieved results for the test query "*wimbledon trophy award*" appears in Figure 2.3.

Evaluation using crowdsourcing

In the known-item scenario used in 2012 and 2013, a single relevant segment was defined by survey participants for each query. The ad-hoc evaluation of the system used in 2014 required a different type of evaluation process. Because of the large amount of results submitted to the evaluation by the task participants and the time-consuming process of watching and judging each video segment, the evaluation was performed using crowdsourcing implemented by the Amazon Mechanical Turk platform. Each participating team was allowed to submit up to

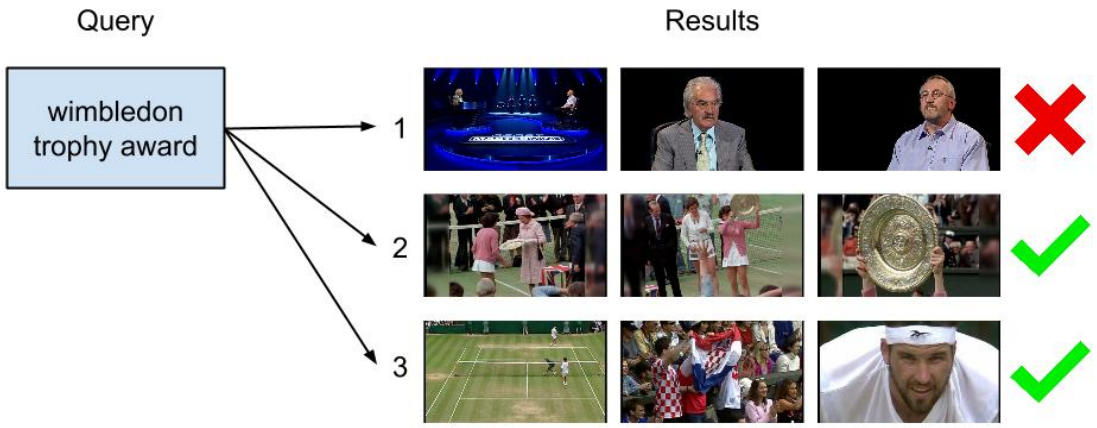


Figure 2.3: Example of the first three results retrieved in the Search sub-task of the 2014 SH Task. Each segment is represented by three keyframes. The first retrieved segment is a false positive, the second and third retrieved segments are correct.

5 runs for each transcript. Pooling was applied to all the submitted runs and the top 10 ranked pooled runs were evaluated.

2.3 Experiments with text-based search

Depending on the type of data, a video retrieval system can make use of appropriate multimedia features such as color, texture, shape, motion, displayed text, and loudness (Patel and Meshram, 2012). A common approach in speech retrieval is to make use of transcripts or subtitles (Larson and Jones, 2012a, p. 239). Spoken terms are then indexed and sometimes combined with additional information such as time of the utterance, weight, or speaker identification. This speech approach can also be combined with the stated multimedia features (Westerveld et al., 2003; Iyengar et al., 2005; Ah-Pine et al., 2015).

Text-based search at the MediaEval Benchmark

A number of retrieval approaches were examined at the MediaEval Benchmark. Various segmentation approaches aimed at improving the quality of the relevant segment retrieval were tested (Eskevich et al., 2012d; Le et al., 2014). Authors also experimented with transcripts and subtitles, augmenting them by using synonyms and conceptually-connected words (Paróczi et al., 2014), stemming, adding stopwords as well as applying different language models (Chiu and Rudnicky, 2014). Simon et al. (2015b) experimented with hierarchical structuring

of topically-focused fragments of videos based on burstiness of word occurrences in the segments (Simon et al., 2015a). Le et al. (2014) proposed a complex system which combined textual search with the use of visual properties. Eskevich and Huet (2015) also experimented with combinations of text and visual concepts, while Racca et al. (2014) experimented with prosodic features.

2.3.1 The baseline system

The setup used in this thesis is based on speech retrieval. ASR systems are first applied to an audio track. Subtitles may be used instead of transcripts, if they are available. The timestamp of the utterance in the subtitles or transcripts is assigned to each word. The timestamp of the utterance is often available in the speech transcripts but it needs to be approximated in the subtitles where only the timestamps of displaying of the full utterance are given. In such cases, it is assumed that each word is equally long and the duration of the utterance is divided into the number of included words.

Afterwards, the IR system is applied to the transcripts or subtitles. The Terrier IR platform (Ounis et al., 2006) is used in all experiments. Terrier is an open source IR platform created at the University of Glasgow. It offers a large number of IR methods and the possibility of including newly implemented methods in Java.

Passage Retrieval is utilized in the setup: all recordings are divided into shorter segments to which a standard retrieval process is applied. Such segmentation enables location of precise relevant segments instead of full relevant recordings and it can also improve the quality of retrieval of full documents (Galuščáková and Pecina, 2014a). The formed segments which may partially overlap serve as documents that are indexed by Terrier. Indexed documents thus consist of all words from subtitles/transcripts lying within the corresponding segment. Finally, IR methods are run on the indexed dataset and relevant segments are retrieved. An overview of the system is shown in Figure 2.4.

2.3.2 Tuning the IR model

Applied IR methods need to be carefully tuned. For the purposes of this work, tuning refers to selection of proper IR methods and setting their parameters. Tuning is particularly important as the IR methods were originally created for retrieving full-length textual documents. In the presented experiments, the database consists of relatively short sub-segments of documents, which may contain errors and specific words provided by the ASR systems. Some words may be incorrectly recognized and the vocabulary may be restricted. In this chapter I present our

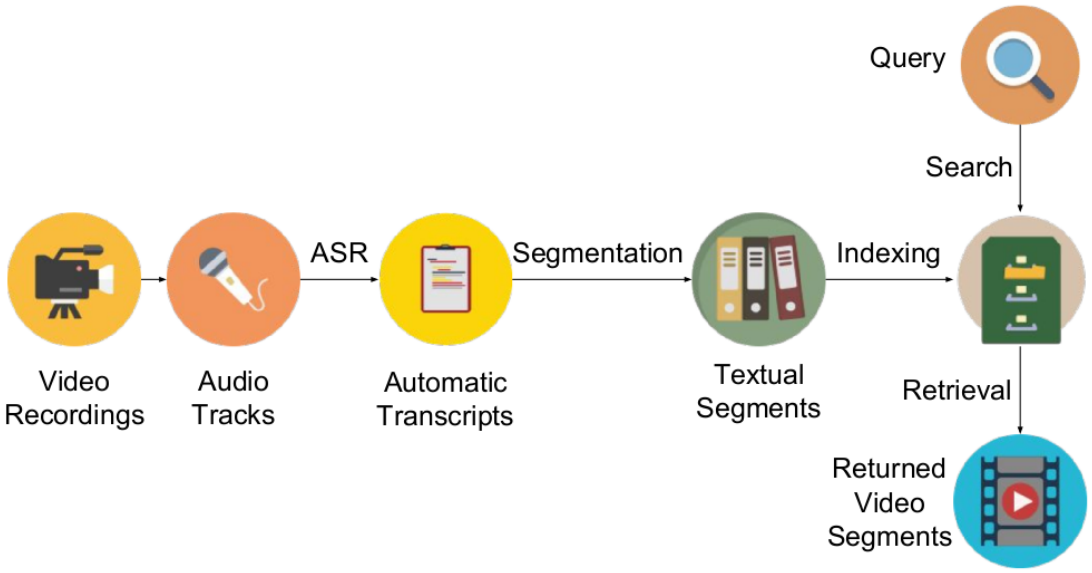


Figure 2.4: The structure of the setup used in the baseline system

experiments with the most frequently used IR models as well as their settings. We performed the tuning experiments presented in this section in the MediaEval 2012 SH Task (Eskevich et al., 2012b). The Blip.tv dataset was used in all of these experiments which were conducted on the automatic transcripts provided by the LIMSI and the LIUM teams.

IR models

The most commonly used IR models are examined in this section: the traditionally used vector-based *TF IDF* model (Manning et al., 2008a, p. 118), the *language model* – specifically the *Hiemstra Language Model* (Hiemstra, 2001), and the *probabilistic model* – specifically the *BM25* (Manning et al., 2008a, p. 232).

TF IDF model definition

The *TF IDF* score is calculated by multiplying the term frequency (TF) value by the inverse document frequency (IDF) value. For document d and term t, the TF value indicates the number of times t occurs in d. The TF value may be logarithmically normalized:

$$TF = 1 + \log f(t, d)$$

where $f(t, d)$ represents the number of occurrences of t in d. Alternatively, diverse normalization methods may be used instead.

The IDF value is calculated as the inverse value of the document frequency, which indicates the number of documents in a dataset which contain t. This value reports how informative the term t is in the dataset. Less informative terms will occur in nearly all documents, while more informative terms will only occur in specific documents. The inverse document frequency is then logarithmically scaled:

$$IDF = \log(1 + \frac{N}{f(t, N)})$$

where N is the number of documents in the dataset and $f(t, N)$ is the number of documents from the dataset which contain t. This value may be further scaled or weighted in different ways.

Each document as well as each query may then be represented as a vector of the TF IDF values calculated for each word in the dataset. Both documents and queries are represented in the vector space defined by the words in the dataset. The distance between the query (q) and each document (d) can thus be calculated using the cosine similarity between these vectors:

$$Similarity(d, q) = \frac{\vec{d}.\vec{q}}{\|\vec{q}\|.\|\vec{d}\|}.$$

Language models

The concept of the Language Model (LM) is that a user searching for a particular document types a query which is similar to that document. Therefore, documents are retrieved according to the probability that the query q was generated by the document $P(q|d)$ (Manning et al., 2008a, p. 237). Language modeling approach assumes that a probabilistic LM M_d exists for each document d in the dataset and and that this document was generated according to its LM (Manning et al., 2008a, p. 237). The *query likelihood* LM is frequently used for this purpose. This model ranks documents according to their likelihood of being relevant to the query $P(d|q)$. Then, according to Bayes' rule:

$$P(d|q) = P(q|d)P(d)/P(q)$$

Since the probability of the query $P(q)$ is the same for each document it may be ignored as it does not influence the final results ordering. The probability of the document $P(d)$ is frequently considered to be uniform and may also be ignored. The query terms are assumed to be generated identically and independently from the document and the unigram model may be used for calculating $P(q|d)$:

$$P(q|d) = \prod_{t \in q} P(t|d)^{n(t,q)}$$

where $n(t, q)$ is the number of occurrences of the term t in the query q (Lalmas, 2011). Here, smoothing needs to be applied to avoid zero values for terms not occurring in the documents. The Hiemstra LM, which is used our experiments presented in this thesis, utilizes the *Jelinek-Mercer smoothing*:

$$P(q = t_1, t_2, ..., t_n|d) = \prod_{t=1}^{n}((1 - \lambda).P(t_i) + \lambda.P(t_i|d))$$

where λ is the importance of the query term.

Probabilistic IR models

In *probabilistic IR models* (Robertson and Walker, 1994), documents are ranked according to their probability of being relevant to a given query. According to the *Probability Ranking Principle* these ranking methods result in the most effective IR systems. One such model, BM25, used in this work, is a well established, probabilistic model. According to BM25, the probability of the relevancy $R(d, q)$ between the document d and the query q is calculated as follows:

$$R(d, q) = \sum_{t \in q}(w_t.\frac{(k_1 + 1).tf(t, d)}{K + tf(t, d)}.\frac{(k_3 + 1).tf(t, q)}{k_3 + tf(t, q)}) + k_2.|q|.\frac{avgdl - dl}{avgdl + dl}$$

and K is calculated as:

$$K = k_1((1 - b) + \frac{b.dl}{avdl})$$

where w_t is the weight of the term t, $tf(t, d)$ and $tf(t, q)$ are frequencies of the term inside of the document and query, respectively, k_1, k_2, k_3 and b are tuning parameters, dl is the document length, and $avgdl$ is the average document length (Lalmas, 2011). The BM25 model is frequently used with Okapi weighting (Robertson et al., 1994). In Terrier, the following parameters are used: $k_1 = 1.2$, $k_2 = 0$, $k_3 = 8$ and $b = 0.75$.

Retrieval model performance comparison

The described models are frequently compared with each other, also in video retrieval. Chiu and Rudnicky (2014) achieved the best results using vector-based TF IDF. Cheng et al. (2015) confirmed this result and remarked that the best performance was obtained by TF IDF for text information without context, while the probabilistic methods performed better when context was considered and indexed documents were longer. The performance of models compared using the Blip.tv dataset is tabulated in Table 2.8.

	LIMSI			LIUM		
	MRRw	mGAP	MASP	MRRw	mGAP	MASP
LM	**0.470**	**0.290**	**0.123**	**0.449**	**0.250**	**0.102**
TF IDF	0.428	0.256	0.103	0.418	0.239	0.087
BM25	0.423	0.251	0.102	0.429	0.238	0.091

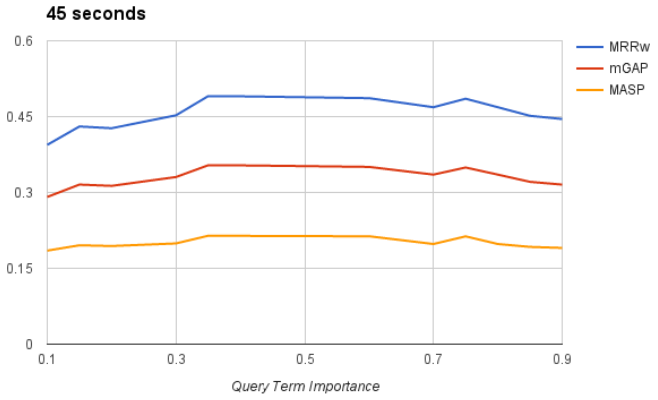
Table 2.8: Retrieval model performance comparison. The resulting scores of all systems are using the default setting, for 90-second-long windows with 30-second overlaps, with stemming, stopwords, metadata and "overlap removal" filtering and both "Title" and "Short Title" fields employed. The best results are in **bold type**.

For both the LIMSI and LIUM transcripts, the LM achieves the highest score. In case of LIUM transcripts, BM25 slightly outperforms the TF IDF model. In the case of LIMSI transcripts, TF IDF is slightly better than BM25, but the difference is minor.

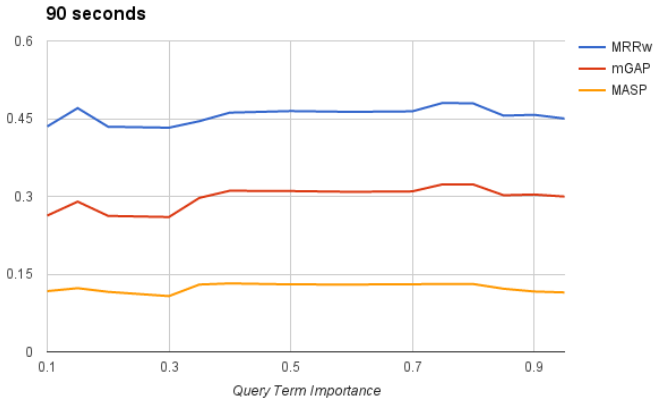
Tuning the LM

The results of Hiemstra LM are highly dependent on the parameter indicating the importance of a query term in a document (Hiemstra, 2001). In the presented experiments, there is a correlation between segment length and the Hiemstra LM parameter, herein designated the "*query term importance*". This behaviour is apparent in Figure 2.5 (all the experiments are performed on the LIMSI transcript).

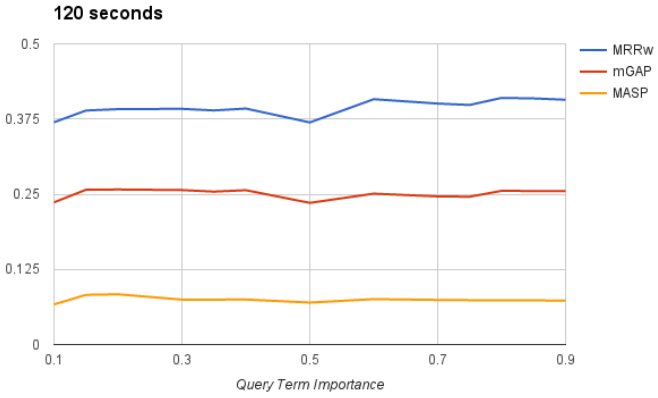
Testing the evaluation measures show that they exhibit different in maximum score values. For 45 second long segments, the highest values for all measures are achieved using the parameter value 0.35. Measure values for 90-second segments achieve maximum for MRRw and mGAP scores using the parameter value 0.75 and for the MASP score with the parameter value 0.4. For 120-second segments, the maximum MRRw score is achieved using 0.8 and the maximum mGAP and MASP scores are achieved using 0.2. In all cases, there is a local maximum of the function at the parameter value of 0.15, then the function breaks around the parameter value 0.35 and the next local optimum occurs around 0.75. Hiemstra (2001) also experimentally determined that a parameter value of 0.15 performs well in general.



(a)



(b)



(c)

Figure 2.5: Behaviour of the Hiemstra LM parameter on LIMSI transcript scores for (a) 45-second-long segments with 15-second overlaps, (b) 90-second-long segments with 30-second overlaps and for (c) 120-second long segments with 30-second overlaps with stemming, stopwords, metadata and "overlap removal" filtering and both "Title" and "Short Title" fields employed.

Stopwords, stemming and full query application

Stopwords filtering and *stemming* are standard pre-processing procedures used in IR. Using a stopword list, most common words having low information value are filtered out. *Stemming* is used to find the root of each word and thus reduce the complexity of the text by removing word suffixes; for example, the words "running" and "run" are stemmed into the same root "run", and the word "ran" is stemmed to the root "ran". This method is especially helpful in the case of highly inflectional languages such as Czech and Slovak. Lemmatization is sometimes used instead of stemming. *Lemmatization* is usually a more complex process which returns a dictionary form of the word (e.g. words "running", "run" and "ran" are all lemmatized to the word "run"). The effect of employing the stopwords filtering and stemming to the Blip.tv dataset is tabulated in the Table 2.9. Procedures available directly in Terrier are applied in the presented experiments: implicit stopword list filtering and Porter stemming (Porter, 1997). Application of stopwords and stemming procedures improve the results; in some cases, almost by a factor of two.

	LIMSI			LIUM		
	MRRw	mGAP	MASP	MRRw	mGAP	MASP
Baseline	0.195	0.131	0.049	0.242	0.155	0.062
Stopwords + Stemming	0.291	**0.211**	**0.079**	0.313	0.164	0.064
Title + Short Title	**0.310**	0.188	0.077	**0.321**	**0.171**	**0.079**

Table 2.9: Applying stopwords filtering and stemming to the LIMSI and LIUM transcripts. In both cases, LM is applied to 90-second-long segments with 30-second overlaps. The best results are in **bold type**.

Only the "Title" field of the query is used in the baseline run displayed in Table 2.9. However, application of the "Short Title" also improves the baseline results. In the case of all evaluation measures with the LIUM transcripts and in the case of the MRRw measure with LIMSI transcripts, the use of the "Short Title" field helps the baseline run even more than applying stopwords and stemming procedures.

Filtering of overlapping results

When segmentation produces overlapping segments, overlapping passages are also present in the retrieved results. Such overlapping retrieved segments lower the comfort of the users of the retrieval system as the same content may be presented several times. Users thus need to filter out duplicate content manually.

As the evaluation scores reflect the user needs, overlapping may also cause a decrease in the MRRw and mGAP scores. The main cause is that if there are numerous irrelevant overlapping segments with high rank in the list of retrieved results, the rank of the relevant segment in the list is then lowered.

Therefore, several strategies to remove overlapping were applied: only the highest ranked segment from each document (*single best*) was used, the segments which partially overlap higher ranked segments were filtered out (*Overlap Removal (ROO)*), see Figure 2.6, and all segments which lay in the vicinity of higher ranked segments were filtered out (*window filtering*), see Figure 2.7.

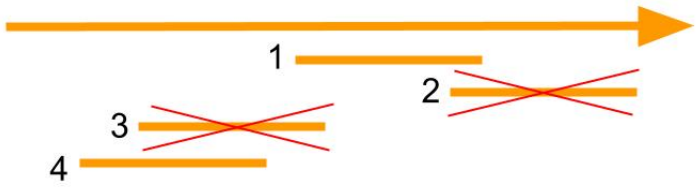


Figure 2.6: Overlap Removal Filtering: segments overlapping other higher ranked segments from the list of retrieved results are removed.

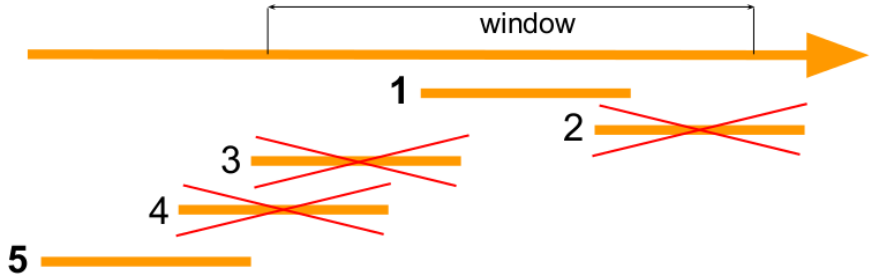


Figure 2.7: Window Filtering: Removal of segments laying in the vicinity of higher ranked segments.

For comparison of these filtering methods see Table 2.10. Pursuant to these results, the hypothesis that overlapping segments in the retrieved results could decrease the overall score is confirmed. The most efficient strategy for filtering the results is to remove all segments which partially overlap with higher ranked segments.

Metadata utilization

All recordings in the Blip.tv dataset are accompanied with detailed information such as title, description and tags provided by authors and sometimes

	MRRw	mGAP	MASP
No Filtering	0.474	0.341	0.208
Overlap Removal	**0.489**	**0.352**	**0.214**
Window Filtering	0.486	0.350	0.212
Single Best	0.469	0.335	0.207

Table 2.10: The effect of filtering the results on the LIMSI transcripts. In all cases, Hiemstra LM with parameter 0.35 is applied to 45-second-long segments with 15-second overlaps, with stemming, stopwords, metadata and both "Title" and "Short Title" fields employed. The best results are in **bold type**.

comments entered by viewers. For each segment, metadata belonging to its associated video is employed. Segments from the same file thus share the same metadata. The transcript of each segment is then concatenated with the relevant metadata. This approach improves performance of all evaluation measures, see Figure 2.8.

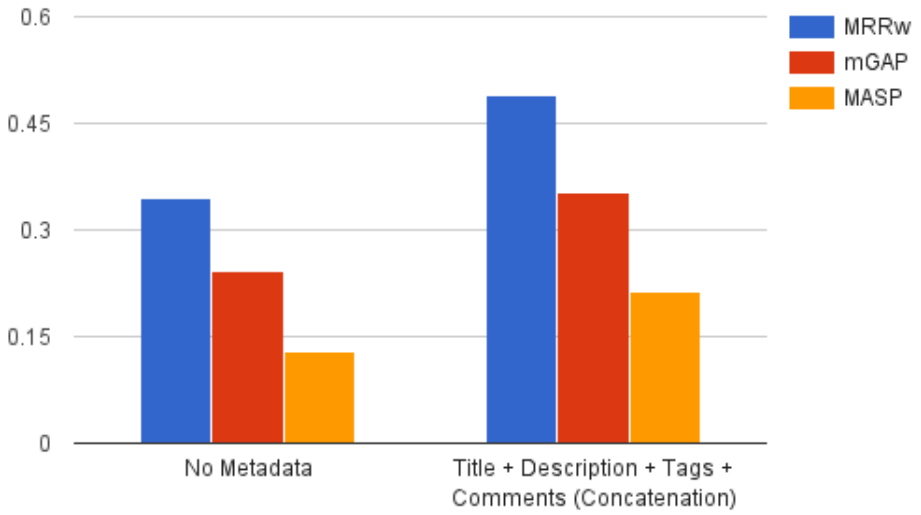


Figure 2.8: Effect of using metadata on the LIMSI transcripts. Hiemstra LM with the parameter 0.35 is applied to 45-second-long segments with 15-second overlaps, with stemming, stopwords, and ROO filtering employed.

More complex approaches of applying the metadata description were examined by Chen et al. (2014) who tuned the fusion weights used in the combi-

nation of transcripts and metadata. They proved that complex determination of weights of the fusion of transcripts and metadata using Linear Discriminant Analysis can also improve search performance. Comparison of this more complex fusion method and simple combination of transcripts and metadata used in our experiments should be further explored.

Impact of the transcripts

In the presented experiments, the LIUM transcripts outperform the LIMSI transcripts in the baseline run but in the tuned runs (stopwords, stemming, metadata, filtering, and short title employed), LIMSI achieve higher scores, see Table 2.11.

	LIMSI			LIUM		
	MRRw	mGAP	MASP	MRRw	mGAP	MASP
Baseline Run	0.195	0.131	0.049	**0.242**	**0.155**	**0.062**
Tuned Run	**0.470**	**0.290**	**0.123**	0.449	0.255	0.102

Table 2.11: Comparison of LIMSI and LIUM scores for baseline and tuned runs. The results of the LM are without parameter tuning, for 90-second-long segments with 30-second overlaps, with stemming, stopwords, metadata and ROO filtering and both "Title" and "Short Title" fields employed. The best results are in **bold type**.

In the case of the LIMSI transcripts, all available word variants are used, and in the case of the LIUM transcripts, the single-best output is used. The average word error rate of relevant passages is 0.317 and 0.404 for the LIMSI and LIUM transcripts, respectively (Eskevich et al., 2013). As LIMSI offers more word variants, the transcripts are more robust than the single-best LIUM transcript. The transcripts also differ in vocabulary. The vocabulary of the LIMSI transcripts is larger but it is mainly due to transcribing into several languages, as previously mentioned in Section 2.1. If only English files from the dataset are used, the size of the vocabulary drops by more than a half, see Table 2.12.

	LIMSI	LIUM
Words Total	13.9 mil	10.3 mil
Words Unique	186 k	87 k
English Words Total	12.6 mil	9.1 mil
English Words Unique	93 k	81 k

Table 2.12: Statistics of Blip.tv test data – English vs. full vocabulary.

Query Expansion and Pseudo-relevance Feedback

Query Expansion and *Relevance Feedback* are methods used to address the problem of small lexical overlap of the query and a relevant passage. Among other things, they provide help in the case when the relevant document contains synonyms of the terms from the query but not identical words.

Relevance Feedback (Manning et al., 2008a, p. 178) improves the retrieval quality by involving users of the IR system into the retrieval. After the IR system returns a set of results, users can mark some of the retrieved documents as relevant or irrelevant. IR thus acquires better knowledge about the user's information need and it can then recalculate the scores of the rest of the documents in the dataset using this knowledge. *Pseudo-relevance Feedback* (Manning et al., 2008a, p. 187) is an automatic adaptation of Relevance Feedback. It is assumed that several top ranked documents, which were retrieved using the original query, are relevant. The query is then automatically adjusted towards these highest ranked documents from the dataset.

In the presented experiments, Pseudo-relevance Feedback increases the MRRw score and decreases the mGAP and MASP scores of the baseline, see Figure 2.9. If Pseudo-relevance Feedback is employed in a case when stemming, stopwords, metadata, ROO filtering and both "Title" and "Short Title" fields are used, all scores drop. The improvement of the Pseudo-relevance Feedback is thus not clear. Moreover, Pseudo-relevance Feedback cannot be expected to increase the quality of the retrieval when query and relevant passage are already expanded by the metadata as in such case Pseudo-relevance Feedback possibly favors irrelevant documents.

Query Expansion is a popular IR technique (Papka and Allen, 1997; Allan, 1995; Xu and Croft, 1996) which enables expansion of a query with new words related to the original query. Queries in the presented experiments were expanded using WordNet (Miller, 1995). WordNet is a large lexical database of English nouns, verbs, adjectives and adverbs. Words are segregated into synsets and semantic relations between words, such as hyperonymy, hyponymy, ISA relations, and meronymy are available. Each query term was expanded by a set of coordinated terms, derived words, hypernyms, hyponyms, and synonyms. Results for different expansion terms are reported in Figure 2.10. The correct sense was not disambiguated and in each case, all possible senses were used. As this approach brought too much additional noise to the retrieval, all strategies decreased the scores. Expansion by *derivation of nouns* produced the smallest decrease.

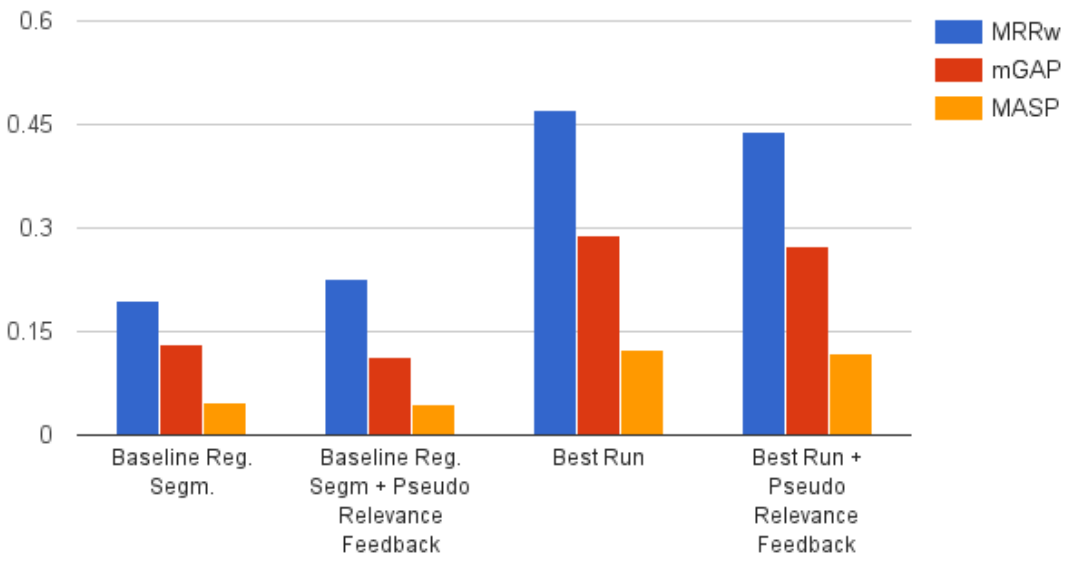


Figure 2.9: Effect of using Pseudo-relevance Feedback on evaluation measure scores of LIMSI transcript retrievals.

The best achieved results

The best result from the text-search experiments was achieved on the LIMSI transcripts using the Hiemstra LM with the parameter 0.35, for 45-second-long segments with 15-second overlaps, with stemming, stopwords, metadata and ROO filtering and both "Title" and "Short Title" fields applied. The results for the best run are displayed in Table 2.13.

	MRRw	mGAP	MASP
Best Run	0.489	0.352	0.214

Table 2.13: Results of the Best Run; achieved using LIMSI transcripts with the Hiemstra LM with parameter 0.35, for 45-second-long segments with 15-second overlaps, with stemming, stopwords, metadata and ROO filtering and both "Title" and "Short Title" fields applied.

Compared to other approaches tested in 2012 at the MediaEval Benchmarking, this setting achieved the highest scores from all evaluation measures, see Figure 2.11 (Eskevich et al., 2012a). Tuning the parameters of the system proved to be extremely important in our case. The biggest improvement in our results was achieved by combining retrieval using transcripts and additional metadata. Segmentation method, especially the segment length, and filtering of the results were also very important.

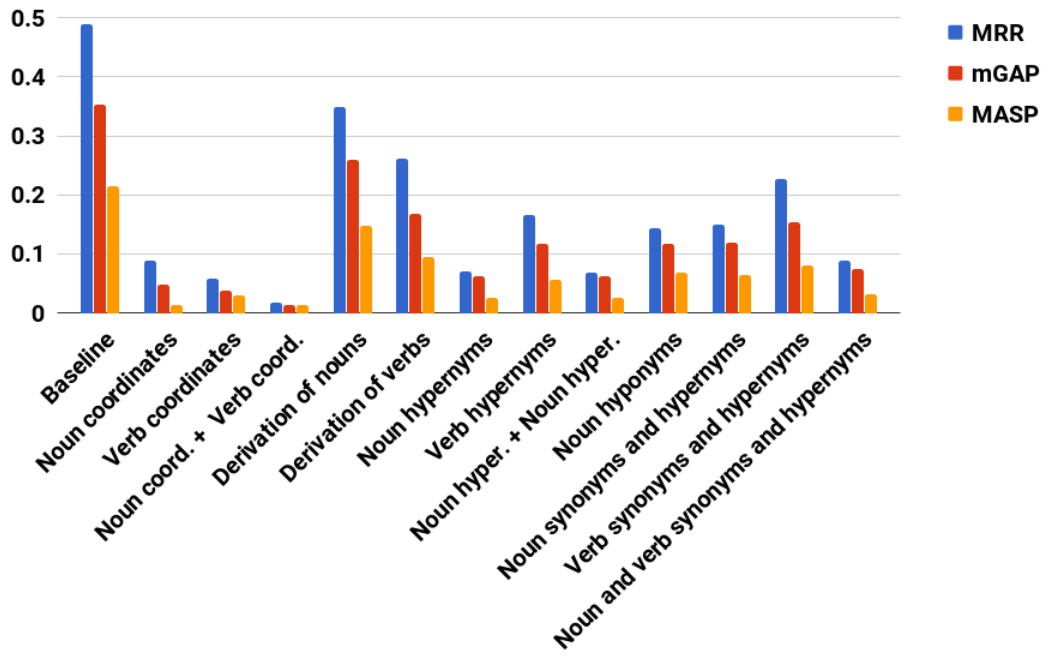


Figure 2.10: The effect of query expansion using WordNet on the LIMSI transcripts. Hiemstra LM with parameter 0.35 is applied to 45-second-long segments with 15-second overlaps, with stemming, stopwords, metadata and ROO filtering and both "Title" and "Short Title" fields employed.

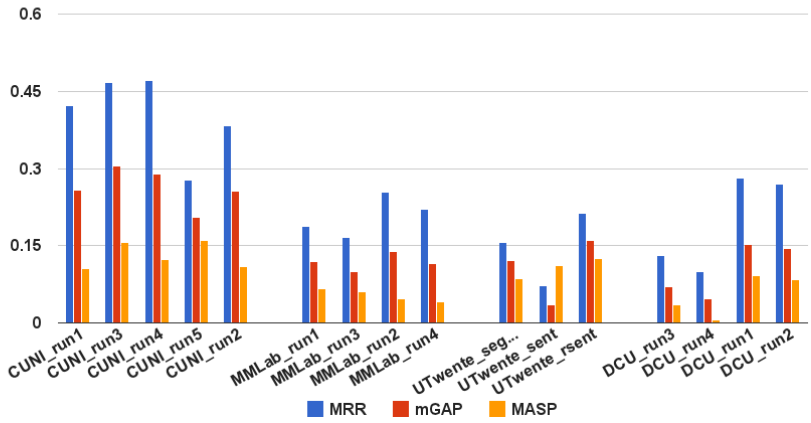


Figure 2.11: Results of the Search sub-task organized at the 2012 SH Task at MediaEval. Our team results are labeled as *CUNI*.

Results for individual queries

In Figure 2.12 and Figure 2.13, the evaluation measure results for each query for both LIMSI and LIUM transcripts are depicted. The queries associated with

the highest MRRw scores are displayed in Table 2.14 and those associated with the lowest MRRw scores in Table 2.15.

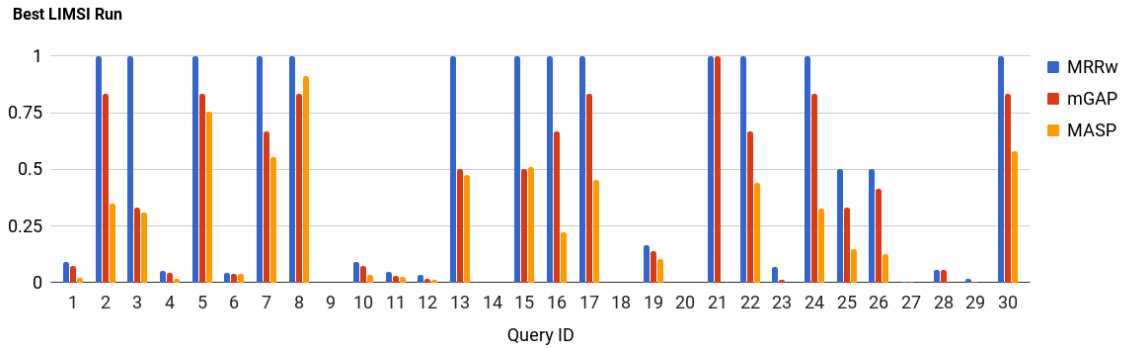


Figure 2.12: Measure results for the Best Run of each query using LIMSI transcripts.

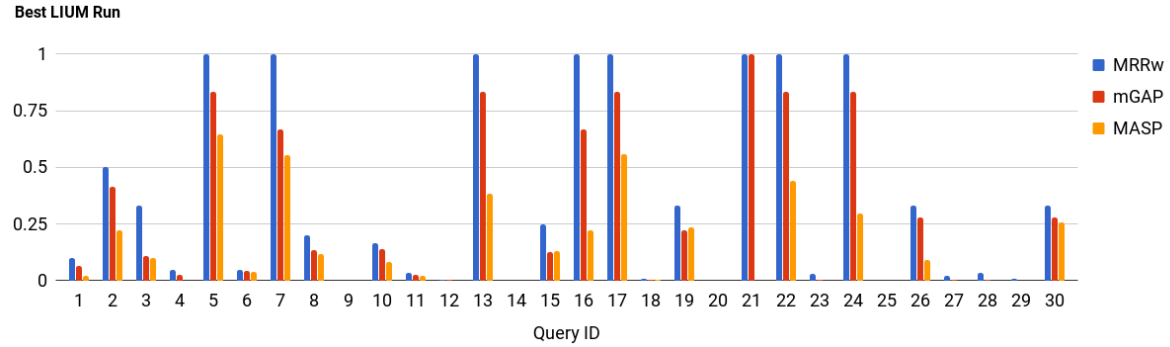


Figure 2.13: Measure results for each query for the Best Run of each query using LIUM transcripts.

Since the LIMSI transcripts error rate is lower than the LIUM transcripts error rate, not surprisingly, search performance using LIMSI transcripts outperforms LIUM transcripts for most of the queries. However, in the case of queries 10 and 19, LIUM transcripts achieved better scores for all evaluation measures and for query 17, the LIUM transcripts outperformed the LIMSI transcripts in the MASP score. Query 21 is also unique: it achieves maximum MRR and mGAP scores, but its MASP score is equal to zero. Generally, queries with high scores often contain specific words and proper names, which help identify the segment of interest. Queries associated with low scores are usually very descriptive. They provide a description of the content of the segment but they do not contain words which can be expected to occur in the relevant segment. These queries are also closer to queries typically used in Question Answering task. Especially, query 20

Id	Query
2	Profit Partner programe talks about growing business faster.
3	Curtis Baylor of Allstate gives a small piece of planning advice for small business using his basic three factors.
5	One of the biggest problems with the EEE PC 900 laptop and how to solve it.
7	Its about an annual Brooklyn Blogfest where bloggers and fans meet each other and have fun.
8	"Hey guys, I thought this was pretty…interesting to listen to. Minus the fact it should be Judaism,and not Judism (sounded like Druidism HAH) I thought his reaction to the news of conversion was pretty funny."
13	Medical Marijuana clinics in California.
15	Its about wrong impressions created by artists on Angels and clarifies the authentic interpretation as per the Bible.
16	California to pass law intended to put an end to domestic violence by outing the abusers in public.
17	What an unusual painting interview
21	Too Big to Fail composed by Austin Launge Lizards
22	Its a Grit TV presentation on Green Party Presidential Candidate.
24	Sending automatic emails whenever you add new content to blogs or web sites.
30	What Would Google Do By Jeff Jarvis

Table 2.14: Queries associated with the highest MRRw score (equal to 1) using the LIMSI transcripts.

Id	Query
9	Its of serious comics on science related subjects.
14	This is the process a comic book goes through before it's released.
18	"This is a video that includes two different poets, both doing readings of their work."
20	"I found this clip simple but very helpful. I couldn't remember how to create a new new pattern, but the steps were pretty simple and easy to follow. Hope it can help you guys out too! Enjoy."

Table 2.15: Queries associated with the lowest MRRw score (equal to 0) using the LIMSI transcripts.

is formulated very freely and it can be hardly expected that any relevant segment will be retrieved for it.

Conclusion

This section summarizes our experiments with text-based retrieval. We provide an extensive study of parameters of a text retrieval system. The baseline system makes use of automatic transcripts of the audio track and segmentation of the recordings allows searching for relevant segments of the videos. Using parameter tuning, the highest achieved results are doubled comparing to the baseline setting and helped to achieve state-of-the-art results. Even though some of the tuning experiments did not improve the results, this kind of research is still needed as this type of experiments are expensive to conduct and no standardized test sets are in existence.

Out of the explored well know retrieval models, LM achieved the best results on both available transcripts. The parameter of the LM was tuned and values 0.15 and 0.75 achieved the best results. Stopwords removal and stemming, which are well applied in IR systems, also improved the results of text based search. Segmentation is extremely important for being able to find the relevant information quickly and 45-second-long segments performed well in our experiments. Filtering of the list of results is needed if the created segments partially overlap, so that duplicate content is removed. Additional metadata available for the documents can be very helpful and can substantially improve performance of the system. However, query expansion using related words adds lots of ambiguity to the queries and thus lowers the performance of the system. Pseudo-relevance feedback can slightly improve the results if no additional metadata is available.

3

Passage Retrieval

Since multimedia recordings tend to be quite long, search engine users may wish to find the exact starting points of relevant passages instead of entire recordings. In *Passage Retrieval*, the recordings are automatically segmented into smaller parts, to which standard retrieval techniques are applied. Various techniques for segmentation of audio-visual recordings are discussed in this chapter. Special attention is focused on machine learning approaches which determine segment boundaries based on various features.

Retrieval of full relevant recordings may be sufficient in the case of short video files which mostly appear on video-sharing websites such as YouTube, where the average length of a video clip is around 4 minutes (Sysomos Inc., 2010). However, in the case of typically much longer TV programmes, retrieval of full documents is not optimal. The average length of a video file in the BBC 2014 SH set is 45 minutes and the longest program, Golf Championship, lasts for more than 10 hours. Many programmes, such as TV news, cover many topics and scanning through such long multi-topic recordings for relevant information is very time-consuming and tedious. Application of Passage Retrieval enables users to find exact relevant segments in a archive of long audio-visual documents and reduces the time required to find the requested information.

Advantages of Passage Retrieval and multimedia segmentation strategies

Passage Retrieval is an IR method which splits texts into smaller units which then function as "mini-documents" in the IR process, thus making the process more precise. Passage Retrieval techniques have also been shown to help "classical" IR in several ways. First, positional information of term occurrences (usually ignored in IR) can be used in indexing and term weighting (Mittendorf and Schäuble, 1994), e.g., by assigning higher weights to terms occurring near the beginning

of documents. Second, Passage Retrieval can improve results of IR for long documents containing a large range of different topics. If a document contains a short relevant passage among many other irrelevant passages, the document is often incorrectly identified as being not relevant. In Passage Retrieval, searched terms must appear within a limited distance which also may subsequently improve the retrieval performance of full documents e.g., (Salton et al., 1993; Kaszkiel and Zobel, 1997). Third, document length normalization (frequently used in IR) can be realized based on the length of the detected segments and not the entire documents. Kaszkiel and Zobel (1997) show that this approach is very useful, especially for similarity evaluation measures which tend to prefer shorter documents (e.g., cosine distance).

Chapter overview

Viable approaches which can be applied to document segmentation are described in this chapter. Special attention is focused on approaches using a document's semantic content based on textual or multimodal data. Experiments with window-based and content-based strategies are then described. Provided experiments were performed using the data from the SH Task and SSSS Task organized during 2012-2014 at the MediaEval Benchmark. The experiments performed best in the comparison with other teams participating in the Search sub-task of the 2014 SH Task, see Figure 3.1[1] (Eskevich et al., 2014a).

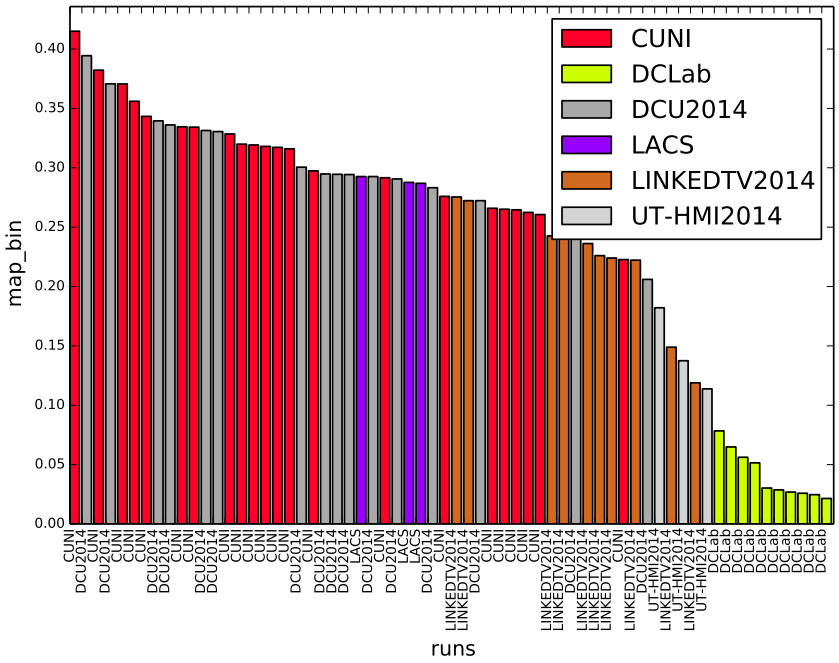
3.1 Approaches to Passage Retrieval

Kaszkiel and Zobel (2001) divide segmentation strategies for Passage Retrieval into three groups: *window-based* (passages are created regularly as overlapping windows of fixed length, measured in terms of words), *structure-based* (defined by the author of the document), and *semantic-based* (corresponds to the real topical structure of documents). Some researchers also use *arbitrary segmentation* in which segments may begin at any arbitrary point in the sentence and last for any length (Liu and Croft, 2002).

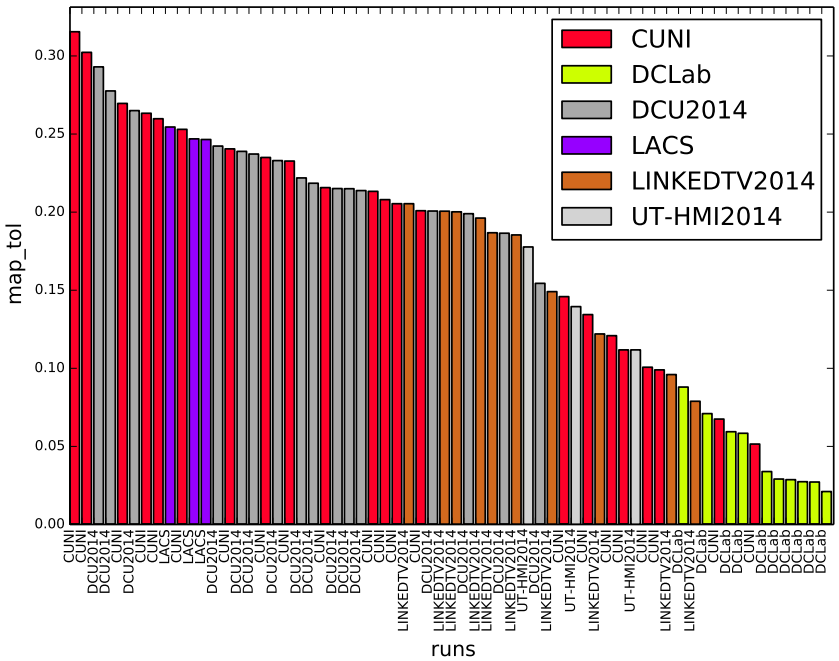
A sliding window approach

Surprisingly, in text retrieval, a majority of researchers, e.g. (Callan, 1994; Kaszkiel and Zobel, 2001; Tiedemann and Mur, 2008), agrees that segmentation using sliding windows and creating overlapping segments of uniform length is the

[1] Figures were provided by the task organizers.



(a) MAP-bin measure



(b) MAP-tol measure

Figure 3.1: A comparison of the team results at the Search sub-task of the 2014 SH Task. Our team results are labeled as *CUNI*

most successful approach to segmentation and its subsequent usage in IR. It is also demonstrated that this approach is sensitive to the window size, which needs to be tuned using training data. For example, Callan (1994) uses a window of about 200 to 250 words, similarly Kaszkiel and Zobel (2001) achieve the best results with 150 to 300 words. Kaszkiel and Zobel (2001) also claim that the segmentation preference depends on the type of query: for short queries and long documents, structure-based segmentation achieves good results, whereas for documents of uniform length, ignoring the document structure is preferred.

A possible explanation for the inferior results related to structure-based segmentation is that it produces segments of highly variable lengths (Kaszkiel and Zobel, 2001). According to Tiedemann and Mur (2008), the actual segmentation method is not critical; it is the length of the segments that is more important in this task. Their semantic-based segmentation based on coreference chains and the TextTiling (Hearst, 1997) algorithm outperforms both: segmentation which is based on paragraphs and sections defined by the author and regular segmentation. They illustrate this improvement in the task of Question Answering.

Passage Retrieval from audio-visual recordings

A window-based approach also achieved good results in audio-visual retrieval experiments by Eskevich et al. (2012d) who compared segmentation techniques of the Rich Speech Retrieval Track participants in the 2011 MediaEval Benchmark. Wartena (2012) compared four segmentation approaches and evaluated the segmentation quality of audio-visual data, achieving the best results on audio-visual recordings with about 20 content words. He concluded that the quality of retrieval is sensitive to segment length. The best results were achieved using a sliding window but segmentation into topically coherent segments proved to be more robust and less sensitive to the predefined average length of the segments. A sliding window achieves better results for longer segments and thus enables a reduction in of the total number of segments.

Other Passage Retrieval-related experiments were performed in the Story Segmentation task in TRECVID 2003 (Smeaton et al., 2003) and 2004 (Smeaton et al., 2004). These tasks were focused on identifying story boundaries in video recordings but the detected boundaries were not subsequently used for IR. In contrast, the goal of the Search task in TRECVID 2003 (Smeaton et al., 2003) was to retrieve shots relevant to given topics but their boundaries were defined as part of the input.

Other applications of Passage Retrieval

Passage Retrieval is used in numerous IR applications, especially in *Question Answering*. Question Answering is a sub-task of IR focused on retrieving exact answers to questions in normal language. To obtain an answer to the assigned question, a relevant passage must first be identified by applying IR to the recordings and marking segments relevant to the question, e.g. (Roberts and Gaizauskas, 2004; Melucci, 1998; Tellex et al., 2003). Then, the answer is mined from the retrieved segment and presented to the user. If no relevant segment is retrieved, the system is unable to return the right answer. Thus IR quality may be considered to be a bottleneck for Question Answering (Tiedemann and Mur, 2008). Question Answering is very sensitive to the length of the relevant segment (Melucci, 1998) which needs to be long enough to contain all relevant information but should not include other, irrelevant, information. In some cases, entire retrieved passages may be considered to be the answer. According to Lin et al. (2003) users even prefer retrievals to be entire passages over exact sentences because passages are embedded in context, which makes the answers seem more trustworthy and simplifies finding answers to related questions.

Another possible application of Passage Retrieval is to facilitate automatic Query Expansion (see Section 2.3.2). It may be beneficial to expand queries by adding related words which occur in the same documents as the query terms. Thus, Query Expansion could be improved by using Passage Retrieval to identify related words in the close vicinity of the original query, i.e. if they occur in the same segment.

3.2 Semantic segmentation

The goal of this work is to improve IR from audio-visual recordings with no or only vaguely specified predefined structure. Segmentation methods originally designed for textual documents are applied to ASR transcripts and combined with other features (e.g., audio, video frames) automatically extracted from the recordings. By using semantic segmentation, the length and makeup of the segments can be controlled. Therefore, semantic segmentation can be an effective method for parceling audio-visual documents for IR. This area of research is relatively new with only a limited number of publications existing on this topic (Eskevich et al., 2012d; Wartena, 2012).

Semantic segmentation denotes segmentation methods which make use of the semantic content of documents. In this section, algorithms for semantic segmentation, which detect passages based on recording content are explored. Segmenta-

tion should be consistent with end results and correspond to expected answers to user queries. Segmentation may also vary according to the type of data archive. For TV news programmes, segments should correspond to individual stories or their parts and thus should be relatively clearly separable. On the other hand, topic boundaries in casual conversation or movies will tend to be more blurred.

In general, segmentation methods can be divided into *similarity-based*, *lexical-chain-based*, and *feature-based* categories (Manning, 1998; Kauchak and Chen, 2005). In the following subsections, these approaches are described with respect to the modality of the input data.

3.2.1 Text segmentation

Most segmentation algorithms which exploit textual information are only based on the measured similarity between potential segments (determined, e.g., by the cosine distance calculated between potential segments). Optimal segments have high intra-similarity (coherence) and low inter-similarity (differ from other segments) (Malioutov and Barzilay, 2006).

Similarity-based algorithms

The two most often used algorithms for semantic segmentation are *TextTiling* (Hearst, 1997) and *C99* (Choi, 2000). Both measure segment similarity by calculating the cosine distance between neighbouring segments. *C99* calculates similarity between all sentence pairs using the cosine measure to create a similarity matrix and identifies regions with high intra-similarities along the diagonal of the matrix. This algorithm uses a graphical algorithm called a *dot plot* (Gibbs and Mcintyre, 1970) which identifies coherent segments as areas with high density of word repetitions. *TextTiling* calculates the similarity for adjacent segments of predefined size and points with the lowest values are designated as boundaries. This algorithm has previously been used also for multimedia retrieval in the Rich Speech Retrieval MediaEval Task in 2011 (Eskevich et al., 2012d)

Lexical chain-based algorithms

Both similarity-based and lexical chain-based algorithms make use of lexical cohesion in topical segments. Lexical chain-based algorithms detect lexically related words based on the fact that the number of related words within one segment is typically higher than the number of related words between adjacent paragraphs. A lexical chain is defined as "a sequence of lexicographically related word occurrences" (Kauchak and Chen, 2005; Stokes et al., 2004). A segment

boundary usually occurs at the point where large numbers of lexical chains begin or end.

Repetition of the lexical items can be detected easily and this approach may be improved by using synonyms and subordinates. Morris and Hirst (1988) determine lexically close words from Roget's thesaurus, Nguyen et al. (2011) further utilize word collocations, Mohri et al. (2010) calculate co-occurrence statistics and Kozima (1993) estimates similarities for pairs of words and uses them to find a sequence of lexical cohesiveness. Ponte and Croft (1997) propose a method for detection of small segments which share few common words. They use Local Content Analysis, which detects the essential concept (bag of words, which describe the topic) of two passages. Thus, passages do not have to contain common words but they need to have similar concepts.

Lexical cohesion is also employed in Bayesian approach (Eisenstein and Barzilay, 2008; Jeong and Titov, 2010). Some researchers use Latent Dirichlet Allocation (LDA) to generate an unsupervised model of the topic and then use Gibbs sampling to estimate this model (Nguyen et al., 2011; Misra et al., 2009). Other approaches are based on Hidden Markov Models (Blei and Moreno, 2001; Mittendorf and Schäuble, 1994).

Feature-based algorithms

Feature-based algorithms make use of machine learning techniques which are applied to various features mined from data. An example of a common feature is a *cue phrase*. Ballantine (2004, p.18) and Hirschberg and Litman (1993) define cue phrases as words and phrases which "serve primarily to indicate document structure or flow, rather than to impart semantic information about the current topic" (e.g., "Good evening", "well", "so"). Thus, they easily indicate the beginning or end of a segment. Beeferman et al. (1999) study the effectiveness of various lexical features and show that the best feature is information on whether a given word was also present up to five words in the past. Other well-performing features include, for example, presence of pronouns and named entities. The most effective features are then used to predict the probability that a topic ends at a given word or sentence and the decision on segment breaks is based on these predictions.

3.2.2 Segmentation in audio-visual recordings

Compared to text-based segmentation approaches, most algorithms for audio-visual recording segmentation are feature-based: they employ supervised machine learning techniques applied to various textual, acoustic, and visual features.

The most common type of segmentation in audio-visual recordings is segmentation into shots. These segmentation methods usually identify changes in the set of sequential frames using, for example, color (Chum et al., 2007; Boreczky and Rowe, 1996), edges (Zabih et al., 1995; Lienhart, 1999), or motion (Boreczky and Rowe, 1996; Courtney, 1997). The average length of a shot is typically relatively small, in recent movies it averages about 2.5 seconds (Miller, 2014), and it rarely corresponds to the full topically coherent scene. More advanced methods need to be used to detect such scenes.

Multimodal features for segmentation

Hsueh and Moore (2007) examine a range of features in a Maximum Entropy classifier and conclude that lexical features (i.e., cue words) are the most effective ones but they need to be combined with audio and visual features to achieve optimal performance. As reported, other well-performing features include conversational features (such as silence, change of speaker activity, and amount of overlapping speech), followed by contextual features (dialogue act type and speaker role), prosodic features (e.g., fundamental frequency and energy level in the audio track), and motion features (detected movements, frontal shots, hand movements).

Tür et al. (2001) combine lexical and prosodic cues. Prosodic cues include energy patterns around segment boundaries, duration features (duration of pauses, duration of final vowels and final rhymes, and their normalized versions), and pitch features (fundamental frequency patterns around the boundary, pitch range). Decision tree and Hidden Markov Models are applied to these features. Similar features are also used by Dielmann and Renals (2005) but they apply them in dynamic Bayesian Networks to solve segmentation of recordings of meetings.

Pye et al. (1998) combine audio segmentation algorithms based on the change in acoustic characteristics and on Kullback-Leibler distance between frames. Their shot segmentation is based on the color histogram of the video. Audio breaks are essential in their work, visual breaks are used to support them. Hauptmann and Witbrock (1998) are especially interested in visual features and they use them to detect commercials. Among scene cuts they also use black frames (which often precede commercials), frame similarity (color histogram similarity and face similarity), and motion information. They also integrate information from captions. Other applicable features count hand gestures, corresponding slides, and notes from meetings, if they are available. Malioutov et al. (2007) have introduced an approach which does not require use of transcripts; they analyse the occurrence of acoustic patterns on the audio track.

Speech recognition impact on segmentation quality

Textual features in audio-visual segmentation need to be acquired using an ASR system. However, the quality level of transcripts usually varies somewhat and thus raises the question of how the quality influences IR. Hsueh and Moore (2007) show that, despite a word recognition error of 39%, none of their systems performs significantly worse on ASR transcripts than on reference transcripts. They offer a possible explanation that the same word is misrecognized in the same manner at various locations in the dataset and thus, the cohesion is not influenced. Utilization of multimodal features may also reduce the impact of the transcript quality and segmentation quality may also be improved by using lattices instead of the single one-best hypothesis of the ASR system: Mohri et al. (2010) show relative improvements of up to 2.3%.

3.2.3 Evaluation of segmentation quality

Segmentation quality can be possibly evaluated using standard *Precision* and *Recall* evaluation measures. *Precision* is calculated as the ratio of cases (e.g. frames, words, sentences) from all marked boundaries in which the segment boundary really occurs. *Recall* is calculated as the ratio of cases from all possible boundaries in which the boundary is marked. But the number of possible boundaries can be huge compared to the number of real segment boundaries, causing unacceptably high Recall values.

Therefore, specific evaluation measures have been proposed to estimate the quality of a segmentation system: P_k (Beeferman et al., 1999) and *WindowDiff* (Pevzner and Hearst, 2002). P_k indicates the probability that two sentences randomly selected from the text are correctly determined to belong to the same or different segments. However, Pevzner and Hearst (2002) found that this measure penalizes "false negatives more heavily than false positives" and "over-penalizes near-misses". Therefore, the *WindowDiff* measure was proposed, based on a modified P_k. In *WindowDiff*, a fixed-length window is incrementally moved through the document and the number of times in which the marked segment boundaries differ from real segment boundaries inside of the window is noted.

In the presented experiments extrinsic evaluation is used. Segmentation quality is not evaluated directly but it is evaluated in practice – by evaluating the applied IR. Methods used for evaluation of IR quality were described in Section 2.2.

3.3 Multimedia Passage Retrieval experiments

Segmentation methods for IR of audio-visual recordings applied to the MediaEval SSSS set, BBC 2013 SH set and BBC 2014 set described in Section 2.1 are examined in this section. The tuned setting which achieved the best results in Section 2.3.2 is further employed in all experiments going forward. The presented system employs the Hiemstra LM with its parameter set to 0.35, stopword removal, and stemming implemented in the Terrier system. The ranked lists of retrieved segments are post-filtered using ROO filtering described in Section 2.3.2.

Passage Retrieval methods are trained and explored on the SSSS Task data and then applied to 2013 SH and to 2014 SH Tasks data. Experiments were first submitted to the 2013 tasks and implemented in the R software environment[2] (R Development Core Team, 2008) and were reimplemented after the task in the Weka framework[3] (Hall et al., 2009) which enabled more proper classification classes bias pre-processing. All 2013 experiments presented in this section were thus conducted after the official evaluation of the tasks. Visual information is also explored in 2014 SH Task and both *query text* and *visual cues* are used to construct the queries. All presented 2014 experiments were submitted to the task and officially evaluated by the task organizer, except the case when visual information was used in segmentation process. For all tasks, the IR system is applied to the ASR transcripts (provided by LIMSI and LIUM in the SH Task and by the University of Edinburgh in the SSSS Task) as well as the manual transcripts (subtitles in the case of the SH Task).

Training queries for SH Task

Since the training set for the SH Task in 2013 only consisted of four queries, I created 30 additional queries and the entire set consisting of 34 queries was used for training. 29 recordings were randomly selected from the archive, then I identified short remarkable passages and and formulated the queries to be used to search for those passages (the short remarkable passages were then considered to be ground-truth). The queries were formulated to imitate the style of the original given queries (e.g., “how to prepare Vietnamese spring rolls”, “Thomas Tallis signature”, and “a difference between a hare and a rabbit”).

[2] `https://www.r-project.org`

[3] `http://www.cs.waikato.ac.nz/ml/weka`

Pre-processing of SSSS Task

The task of searching all segments in the similarity set is converted into the task of retrieving all segments similar to single given query segment. Each query segment is specified by the timestamps of its beginning and end. The actual queries are then constructed by including all words lying within the boundaries of the query segments. For each similarity set and for each segment in the similarity set, this segment is considered as a query and the remaining segments in the similarity set as possible ground-truth points. Thus, the total number of queries is equal to the number of all segments in all similarity sets. Since both transcripts are given in separate tracks for each speaker, these tracks are merged into a single track. In the human transcripts, sentences from both transcripts are sorted according to their beginnings to acquire a single sequential transcript as well as the speakers' segments given in the ASR transcripts. While in the ASR transcripts the exact playback time is given for each word, in the human transcripts such information is available only at the sentence level and therefore playback time of a specific word needs to be approximated by assuming equal duration of words in a sentence.

Query formulation in the SSSS Task

Query segments in the SSSS Tasks are specified by their beginning and ending times. The queries are constructed by including all words lying within the boundaries of the query segment in both tracks. It can be assumed that the vicinity of the query segment also contains relevant information which can further improve the quality of the retrieval. However, the experiments with expanding queries by adding words appearing in the vicinity of the query segment (allowing ± 5, ± 10, ± 15, ± 20, ± 30, and ± 60 seconds) yielded no improvement in the results. Attempts were also made to generate the queries from both human and ASR transcripts and apply them to searching both types of transcripts. The queries created from the human transcripts achieved higher scores when applied to both the human and ASR transcripts, therefore they were used in the presented experiments. Employing human transcripts also simulates a real world IR system setting in a close future, since the quality of the automatic transcripts is improving rapidly and it reached human parity recently (Xiong et al., 2016),

3.3.1 Baseline settings and post-filtering the results

The main baseline runs are performed with no segmentation, i.e., each recording contains only one segment spanning the entire length of the recording. The

baseline scores for the SSSS Task are given in Table 3.1 and for the 2013 SH Task, they are included in Table 3.4 (row 1) together with other results.

Segmentation	Filtering	Manual transcripts			ASR transcripts		
		MRR	MRRw	mGAP	MRR	MRRw	mGAP
None	ROO	0.565	0.122	0.012	0.565	0.144	0.012
None	None	0.879	0.315	0.029	0.858	0.333	0.027
Manual	ROO	0.897	0.671	0.277	0.885	0.669	0.247

Table 3.1: Baseline scores for the SSSS Task.

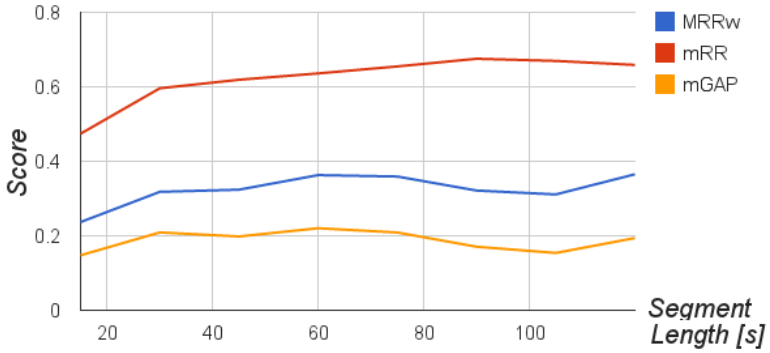
For the SSSS Task, the main baseline (Table 3.1, row 1) is compared to other results: Row 2 refers to experiments without post-filtering (overlapping segments are preserved). This strategy outperforms the baseline experiment with post-filtering since in this case the recordings overlapping the query segments are completely removed from the retrieval results. Row 3 refers to experiments where the IR system is applied to the manually predefined segments. This can be viewed as gold-standard segmentation and the associated scores as a theoretical maximum which could be achieved with the IR system if the segmentation was at an optimum. As expected, the largest room for improvement can be seen in MRRw and mGAP, which take into account the exact starting points of relevant segments (cf. Table 3.1, Rows 1 and 3): The slight increase of the MRR score (from 0.879 to 0.897 for the manual transcripts and from 0.858 to 0.885 for the ASR transcripts) shows that applying segmentation can also improve retrieval of full recordings.

Window-based segmentation

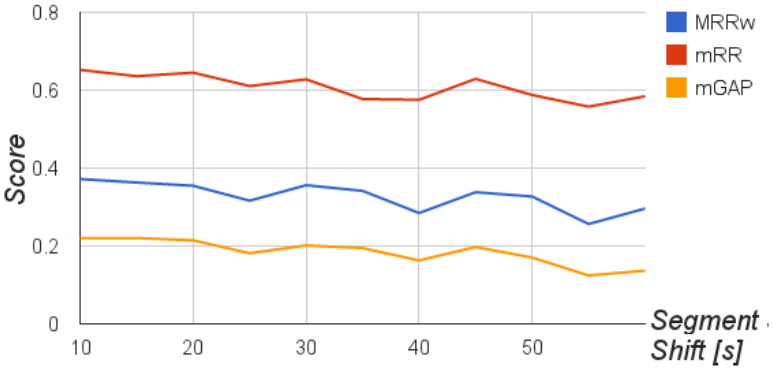
In this approach, sliding windows of various time durations and overlaps are investigated. In contrast to previous approaches where window size is based on the number of words contained within (Wartena, 2012; Kaszkiel and Zobel, 2001; Callan, 1994), the presented strategy uses time duration to size windows. A window having a particular time duration is slid through the transcripts – where it generates new segments of the given length which are sequentially shifted by a particular time distance. The effect of varying the duration of the window (i.e., segment length) appears in Figure 3.2a and the effect of changing the window shift (i.e., segment overlap) appears in Figure 3.2b.

3.3.2 Segmentation strategies

In this and the following sections, the effects of segmentation on retrieval quality are analyzed. *Window-based* segmentation, *TextTiling* segmentation and *feature-based* segmentation using a Machine Learning (ML) approach based on decision trees are explored.



(a) Evaluation scores vs. the length of segments. The shift set to 15 sec.



(b) Evaluation scores vs. the length of segment shift. The segment length set to 60 sec.

Figure 3.2: Evaluation scores vs. the length of segments and the length of segment shift applied in regular segmentation to subtitles in the SH Task.

In Figure 3.2a, the shapes of the MRRw and mGAP curves are similar but slightly differ from the shape of the MRR curve. It is not surprising since MRRw and mGAP are both sensitive to the precise timing information. The highest MRRw and mGAP scores are achieved for 60-second segments and the best MRR score is obtained using about 100-second long segments. Figure 3.2b shows that increasing the segment shift amount consistently degrades the results and the optimal segment shift increment is about 10 seconds. Segments shorter than 10 seconds were not tested since in this case, the number of created segments would be too large and thus impractical for large archives.

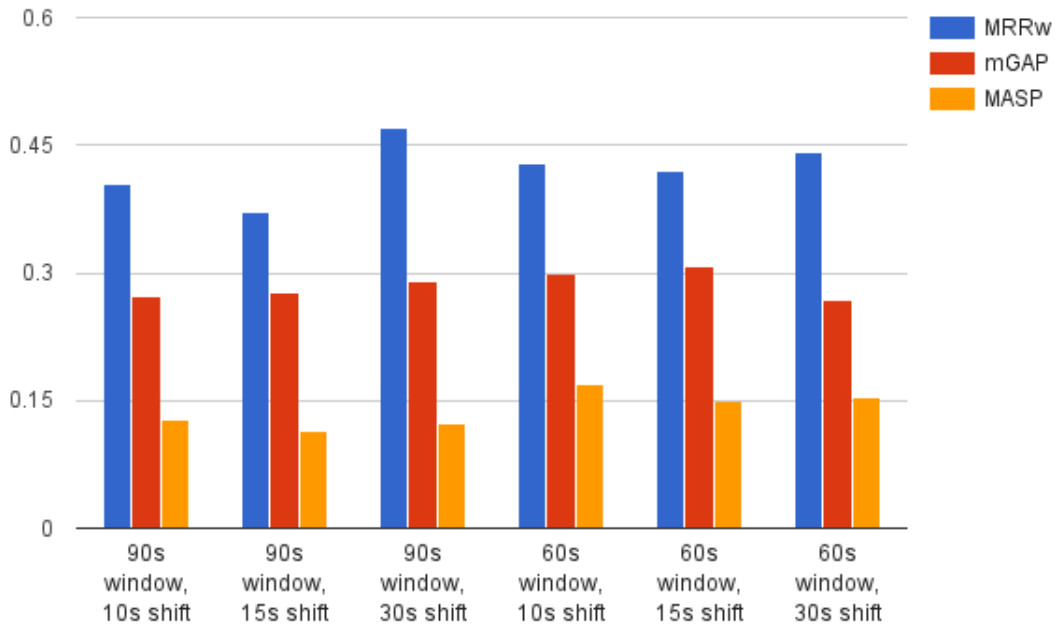


Figure 3.3: The effect of overlapping on the LIMSI transcripts for Hiemstra LM with 60- and 90- second-long windows with various overlaps, stemming, stopwords, metadata and ROO filtering and both "Title" and "Short Title" fields employed.

In the following SH Task experiments, 60 second segments with 10-second shifts are used since this combination achieves the second highest mGAP score and the highest MASP score in the mutual tuning of segment length and segment shift, see Figure 3.3. In the case of window-based segmentation in the SSSS Task, the recordings are divided into equally long segments of 50 seconds each (which is approximately equal to the average segment length in the dataset). The segment shift (or overlap) is also uniformly set to 25 seconds.

TextTiling algorithm

In the second approach, semantic segmentation by employing TextTiling algorithm is studied. The TextTiling algorithm divides input text into semantically coherent segments based on vocabulary usage. The TextTiling algorithm is applied with settings set to correspond to regular segmentation with 90-second-long windows (one segment consists of 9 sentences, containing about 27 words on average). Despite this, the segments created by the TextTiling algorithm are relatively short and therefore it outperforms window-based segmentation in evaluation measures sensitive to short-length segments (MRRw with a 10 second

window and MASP measures). Even with these promising results, TextTiling did not prove to outperform window-based segmentation in subsequent experiments. Therefore, it was not applied to the 2013 data.

3.3.3 Feature-based segmentation

In this approach, segment boundaries (beginnings and ends) are identified using J48 decision trees (Quinlan, 1993). J48 (also called C4.5) is an algorithm for building decision trees based on information gain of individual features available in the training data. The final decision tree is pruned to better avoid over-fitting problem. Unlike some machine learning algorithms, decision trees are easy to interpret and they provide a good overview of the importance of the features. Decision trees are implemented in the Weka framework and trained on the training data available for the SSSS Task containing manually marked segment boundaries.

It is possible to consider this problem to be a binary classification. For each word in the transcripts, it can be predicted whether a segment boundary occurs immediately after it or not. In such a situation, the distribution of the two classes (*segment boundary* vs. *segment continuation*) is highly unbalanced since more words appear inside segments than at their boundaries. Therefore, before training the model, the training data need to be re-sampled to change the class ratio (bias). Ratio values vary from 0.1 to 0.8, by taking as many instances of the *segment boundary* class as possible and as many instances of the *segment continuation* class as needed to achieve a particular ratio. The segmentation is trained on 66% of all examples randomly selected from the training data available for the SSSS Task. The remaining training set is used for segmentation tuning.

Segment formation using detected boundaries

After identification of probable segment boundaries there are several possibilities for formulating the definitive segments. A comparison of three tested strategies for segment formulation is depicted in Figure 3.4.

In the **first experiment**, it is assumed that each word in the transcripts belongs to a single segment; thus, the segments do not overlap. Two variants of this approach are used. First, possible segment beginnings are identified and it is assumed that the previous segment ends at this beginning (further denoted as Beginning = ML and End = "–"). Then, in a similar fashion, possible segment ends are identified and it is assumed that a new segment begins immediately after the detected segment end (denoted as Beginning = "–" and End = ML).

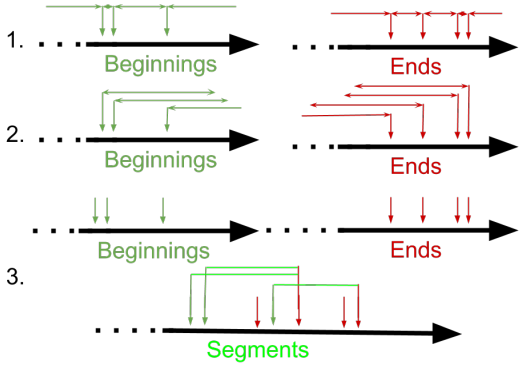


Figure 3.4: Creation of segments using identified probable segment beginnings and ends.

In the **second experiment**, the same process to detect segment beginnings is applied but it is assumed that each segment is 50 seconds long (denoted as Beginning = ML, End = B+50). The segment length was estimated as the average length of manually detected segments in the training data of the SSSS Task. Similarly, the ML model is used to predict segment ends and set the beginnings automatically (denoted as Beginning = E-50, End = ML). In this scenario, the segments can overlap.

In the **third experiment**, all possible segment beginnings and all possible segment ends are first identified. Then, for each possible beginning, the segment end (from the set of identified possible segment ends) which lies closest to 50 seconds from the beginning (Beginning = ML and End = ML) is identified.

Features description

The J48 classification model makes use of lexical and prosodic features and exploits them in boundaries detection decision process. Information about following features is used in the decision trees: cue words and tags, letter cases, length of silence before a word, division given in transcripts (e.g., speech segments defined in the LIMSI transcripts), and the output of the TextTiling algorithm. All the features are binary, indicating whether they appear or not; the length of silence is measured as the difference between timestamps of two adjacent words and is quantized into 15 equal length buckets representing durations from 100ms to 1500ms (in 100ms increments) and corresponding (binary) features indicating whether the length is longer than the corresponding bucket's value or not. A

TextTiling feature indicates whether or not a segment boundary is detected by this tool after the current word.

Cue words

Cue words were identified independently for beginnings and ends by automatic analysis of the training data. Training data were lowercased, punctuation was removed and two sets of words were extracted: *words which frequently appear at segment boundaries* and those *which are the most informative for a segment boundary* (the mutual information between these words and segment boundary is high). In addition, another set of words which did not occur in the training data but are thought to *frequently appear at the boundary* were manually defined.

In addition to cue word unigrams, features for cue words and tag n-grams (unigrams, bigrams and trigrams) appearing at a segment boundary in the training data were used. Tagging was performed by the Featurama tagger (Spousta, 2013). For each type of cue feature (word n-grams, tag n-grams, frequent words, informative words, and defined words for either beginning or end), there is an additional feature indicating whether at least one occurrence of the particular feature type occurs.

The most informative features

The most informative features determined by the performed analysis are divisions defined in the transcripts, the length of silence (especially if it is longer than 300ms, 400ms, 500ms, and 600ms), the output of the TextTiling algorithm, and n-grams of words and tags (especially the features indicating that at least one item of a set of words or tags is present). For example, for segment beginnings, the word n-grams “if”, “I’m”, “especially”, “the”, “are you”, “you have”, and the tag trigram “VBP PRP VBG” (a non-3rd person singular present verb followed by a personal pronoun followed by present participle or gerund – e.g. “are you going”) are highly informative. The letter case feature seems to be informative for segment beginnings. For segment ends, highly informative words are “good”, “the”, “interesting” and “lot” (the article “the” appears in the list of the ending n-grams even though it cannot stand at the end of a sentence, and is probably included due to the approximation of word timing).

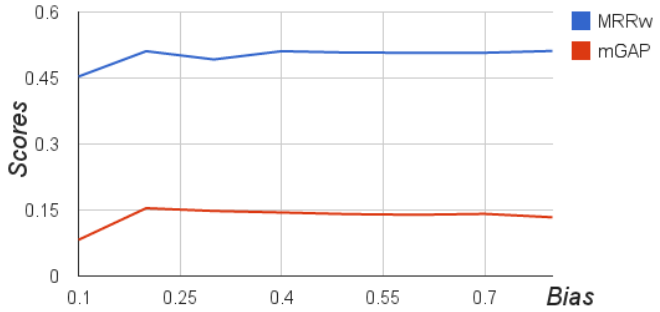


Figure 3.5: The mGAP and MRRw scores vs. class bias in the re-sampled SSSS training set, used for detection of ends of overlapping segments applied to the manual transcripts.

Segmentation tuning

Segmentation was trained on 66% of the SSSS training set. The remaining 34% of the SSSS training set is further denoted as *development set* and it was used for tuning parameters of the segmentation model.

The J48 parameters are set as follows: the confidence factor is set to 0.25 by default for all experiments; the minimum number of instances per leaf is tuned on the development set for each experiment independently and varies from 2 to 250. The final tuning parameter is the class bias in the re-sampled training set (*segment boundary* vs. *segment continuation*), which consequently affects the number of detected segment boundaries and also the retrieval quality. Figure 3.5 shows the effect of this parameter on the retrieval performance measured on the SSSS training set by mGAP and MRRw.

Setting the Weka parameter for class bias to 0.1 leads to around 7% of the words in the development set detected as segment boundaries with mGAP equal to 0.082. Increasing the bias to 0.2 improves mGAP to 0.154 (around 12% of the words were marked as boundaries). Further bias increases cause mGAP to slowly decrease. The MRRw score increases from 0.453 (for the bias equal to 0.1) to 0.511 (for the bias equal 0.2), then it remains almost constant for further bias increases, although there is a slight drop for the bias equal to 0.3.

Segmentation models

Two J48 models were trained and tuned; one to detect segment beginnings and one to detect segment ends. The classification models were trained and tuned on the human transcripts and they then were also applied to the ASR transcripts. Since the number of training queries available in the 2013 SH Search sub-task

Beginning	End	MRR	MRRw	mGAP	#Seg	Len [sec]
–	–	0.565	0.122	0.012	6	719.3
Reg	Reg	**0.858**	0.655	0.233	146	48.4
ML	–	**0.845**	0.626	0.231	1067	7.1
–	ML	**0.858**	0.613	0.164	82	64.2
ML	B+50	**0.859**	**0.690**	**0.255**	1107	47.8
E-50	ML	**0.865**	**0.677**	**0.247**	690	47.9
ML	ML	**0.844**	0.630	0.216	1425	46.1

Table 3.2: Comparison of regular window-based segmentation to several types of feature-based segmentations applied in the SSSS Task for *manual transcripts*.

is very small (even after the our additional queries are included), the SSSS-trained model was applied in the SH Task as well. This allows examination of the possibility of creating a universal model for feature-based segmentation, at the same time, however, it also raises several potential problems. For example: the sets of cue words collected from student dialogues may differ from the cue words used in TV programmes, different ASR systems may prefer different vocabularies, and finally, the silence between words in dialogues may have different distributions compared to the silence between words in TV programmes.

3.3.4 Comparison of segmentation approaches

The complete performance results for the SSSS Task are displayed in Table 3.2[4] (manual transcripts) and Table 3.3[4] (automatic transcripts). In general, the best results are obtained by feature-based segmentation using overlapping segments (rows 5-6) which outperform regular window-based segmentation applied to both types of transcripts evaluated by all three measures, except the MRR score on the ASR transcripts (but this measure is not sensitive to precise starting points of retrieved segments). In terms of MRRw obtained on the manual transcripts, the feature-based segmentation even outperforms the gold-standard segmentation (see Table 3.1). The results of feature-based segmentation using non-overlapping segments (rows 3-4) and feature-based segmentation explicitly detecting beginnings and ends (row 7) are consistently worse than those obtained by regular segmentation.

The 2013 SH Task results for subtitles, LIMSI and LIUM transcripts are displayed in Table 3.4[4], Table 3.5[4], and Table 3.6[4] respectively. Results are not as consistent as in the SSSS Task and differ depending on the type of tran-

[4] Best results and insignificantly lower results for each transcript are in bold type.

Beginning	End	MRR	MRRw	mGAP	#Seg	Len [sec]
–	–	0.565	0.144	0.012	6	680.0
Reg	Reg	**0.834**	**0.615**	0.202	166	48.3
ML	–	0.785	0.538	0.197	1659	3.5
–	ML	**0.809**	0.526	0.131	60	90.7
ML	B+50	**0.818**	**0.623**	0.217	1933	48.5
E-50	ML	**0.820**	**0.616**	**0.226**	964	48.1
ML	ML	0.779	0.538	0.153	2429	68.1

Table 3.3: Comparison of regular window-based segmentation to several types of feature-based segmentation applied in the SSSS Task for *automatic transcripts*.

Beginning	End	MRR	MRRw	mGAP	#Seg	Len [sec]
–	–	**0.656**	0.052	0.027	2 k	2531.6
Reg	Reg	**0.671**	**0.388**	**0.245**	234 k	49.5
ML	–	0.549	0.117	0.060	3125 k	2.3
–	ML	0.607	**0.310**	0.192	280 k	29.0
ML	B+50	0.685	**0.412**	**0.272**	5820 k	49.6
E-50	ML	**0.715**	**0.428**	**0.298**	2580 k	49.6
ML	ML	0.626	**0.392**	**0.229**	5659 k	20.2

Table 3.4: Comparison of regular window-based segmentation to several types of feature-based segmentation applied in the 2013 SH Task for *subtitles*.

scripts. The feature-based approaches creating overlapping segments (row 5-6) are effective especially when applied to subtitles, where they outperform regular segmentation for all evaluation measures. However, for both ASR transcripts, regular segmentation is outperformed when measured by MRR.

Segment counts and their average length differ substantially depending on the transcripts, even with the same segmentation approach. This is probably caused by the difference between the two types of transcripts it is applied to and the fact that the model was trained on data intended for a different task.

In general, feature-based segmentation applied in the two tasks outperforms regular segmentation, which is claimed to be a very effective approach in the previous experiments (Callan, 1994; Kaszkiel and Zobel, 2001; Tiedemann and Mur, 2008). The best approach in the SSSS Task is feature-based segmentation into overlapping segments of regular length. In terms of evaluation measures sensitive to exact segment starting points, the improvement is statistically significant on the manual (MRRw and mGAP measures) and ASR (mGAP measure) transcripts used in the SSSS Task.

Beginning	End	MRR	MRRw	mGAP	#Seg	Len [sec]
–	–	**0.553**	0.052	0.029	2 k	2589.6
Reg	Reg	0.503	**0.299**	**0.172**	242 k	49.5
ML	–	0.455	0.163	**0.119**	664 k	15.0
–	ML	**0.558**	0.180	0.102	20 k	293.5
ML	B+50	**0.484**	**0.276**	**0.165**	748 k	49.6
E-50	ML	**0.468**	**0.256**	**0.159**	435 k	49.5
ML	ML	0.510	**0.250**	**0.141**	931 k	295.6

Table 3.5: Comparison of regular window-based segmentation to several types of feature-based segmentation applied in the 2013 SH Task with the *LIMSI transcripts*.

Beginning	End	MRR	MRRw	mGAP	#Seg	Len [sec]
–	–	**0.566**	0.050	0.028	2 k	2526.5
Reg	Reg	**0.535**	**0.275**	**0.169**	235 k	49.5
ML	–	**0.556**	**0.246**	**0.134**	173 k	63.5
–	ML	**0.561**	0.121	0.051	9 k	854.9
ML	B+50	0.424	**0.180**	0.095	171 k	49.3
E-50	ML	**0.436**	**0.191**	0.087	74 k	48.3
ML	ML	**0.501**	**0.201**	**0.123**	372 k	665.4

Table 3.6: Comparison of regular window-based segmentation to several types of feature-based segmentation applied in the 2013 SH Task with the *LIUM transcripts*.

In the SH Task, the results are not as conclusive, mostly because of the fact that the segmentation model was trained on data for the SSSS Task which differ in many aspects. Still, some of the results, especially results achieved with subtitles, are encouraging and confirm the usefulness of this tested approach.

3.3.5 2014 SH Task experiments

The presented segmentation methods were further explored in the 2014 Search and Hyperlinking (SH) Task. Since the BBC 2013 SH set was used as a training set in the 2014 SH Task, segmentation models described in the previous section were also tested on BBC 2014 SH test set. Compared with previous experiments, one additional automatic transcript provided by NST-Sheffield was available and new evaluation measures were proposed and used. Parameters which achieved the best results in Section 3.3.2 were used (60-second segments with 10-second shifts)

with regular segmentation. In the case of the feature-based ML segmentation model, only a model for detection of segment ends was employed and 50- and 120-second long, possibly overlapping, segments were used.

Additionally, visual information was included in the segment end decision. Since data used in training the segmentation model provided in the SSSS Task are very static, they are not suitable for training employment of visual information in the segmentation process. Recorded interviews typically consists of relatively static image of two people sitting and talking and thus cannot be expected to provide enough information for such training. Therefore, the visual similarity between each two adjacent keyframes was calculated (visual similarity is thoroughly described in Section 4.3.1) on the BBC 2013 SH set and BBC 2014 SH test set. Threshold indicating whether the segment boundary occurs for a given similarity score was tuned on the BBC 2013 SH set. Then, the segment boundary in the test data was only permitted to occur if the similarity between the neighbouring keyframes in adjacent segments was smaller than a given similarity threshold indicator.

Transcript	Metadata	Filtering	Segmentation	Len	MAP-bin	MAP-tol
Subtitles	No	ROO	Reg	60s	0.319	**0.316**
Subtitles	Yes	ROO	Reg	60s	0.343	0.302
Subtitles	Yes	None	Reg	60s	**0.415**	0.146
Subtitles	Yes	ROO	ML	120s	0.261	0.216
Subtitles	Yes	None	ML	120s	0.317	0.052
Subtitles	Yes	ROO	ML	50s	0.320	0.235
Subtitles	Yes	ROO	ML+Visual	50s	0.329	0.253
LIMSI	No	ROO	Reg	60s	0.292	0.263
LIMSI	Yes	ROO	Reg	60s	0.316	**0.270**
LIMSI	Yes	None	Reg	60s	**0.382**	0.134
LIMSI	Yes	None	ML	120s	0.371	0.101
LIUM	No	ROO	Reg	60s	0.223	0.208
LIUM	Yes	ROO	Reg	60s	0.265	**0.233**
LIUM	Yes	None	Reg	60s	0.318	0.112
LIUM	Yes	None	ML	120s	**0.335**	0.099
NST-Sheffield	No	ROO	Reg	60s	0.265	0.241
NST-Sheffield	Yes	ROO	Reg	60s	0.297	**0.260**
NST-Sheffield	Yes	None	Reg	60s	**0.356**	0.121
NST-Sheffield	Yes	None	ML	120s	0.334	0.068

Table 3.7: Results of the Search sub-task of the 2014 SH Task for different transcripts, metadata employment, filtering of the retrieved segments, segmentation type, and segment length.

Results of the Search sub-task of the 2014 SH Task are reported in Table 3.7[4]. The best results were achieved for subtitles and fixed length segments. The highest MAP-bin score was achieved when metadata were utilized, and the returned overlapping segments were preserved. The best MAP-tol score was achieved for the baseline setup; no metadata were utilized and overlapping segments were removed from the list of the retrieved segments. ML-based segmentation improved only the MAP-bin score when it was applied to the LIUM transcripts. For ML-based segmentation, the 50-second long segments outperformed the 120-second long segments. The visual information slightly increased the segmentation performance.

3.3.6 Conclusion

This section concludes our experiments with employing passage retrieval in the text-based search. Our experiments confirm high impact of passage retrieval in multimedia retrieval setting. We also confirm that passage retrieval is not only essential for retrieving relevant segment of the document but it can also improve quality of full document retrieval. We further explore how does segmentation type and segment length influence quality of the retrieval. We specifically focus on regular segmentation in which segments of unified window size are created and on semantic feature-based segmentation which utilizes different multimodal features. The length of the window in the regular segmentation needs to be properly tuned for the dataset and for the specific application but the window of about one minute performed well in general in our experiments. We proved that regular segmentation, which traditionally achieves the highest performance, can be outperformed by well tuned feature-based segmentation. We provide an overview of the most effective features used in the feature based segmentation. We mainly use lexical and prosodic features such as cue words and length of the silence but the performance can be further improved by incorporating visual information. Since the experiments were conducted in three MediaEval tasks in 2013 and 2014, we provide our studies on different datasets and thus verify that this approach performs well even if train data substantially differ from test data. The described approach achieved the highest results in the Search sub-task of the 2014 SH Task.

4

Video hyperlinking

Text-based search described in the previous chapters requires users to express their information needs using a limited number of words. This can be restrictive, especially in multimedia retrieval where users wish to find more information topics discussed, displayed objects or background music. In such cases, the query formulation may be challenging since users may do not know the precise name of the person or object of their interests and typing the query may be considered time consuming. Therefore, a *query-by-example* concept is often used in multimedia retrieval (Rüger, 2010). This concept is especially suitable for image retrieval, whereby users input a query image (Flickner et al., 1995; Shapiro and Stockman, 2001) or even draw or sketch it (Kato et al., 1992; Xiao et al., 2015).

Query-by-example is also effective in *video hyperlinking*, in which segments of the video can form queries. This method thus enables navigation using links between related segments. Ordelman et al. (2015b) define the typical use case of hyperlinking as a "navigation through large quantities of locally archived or distributed video content via a link structure at the level of media fragments". Navigation via links is a well-known concept which is especially suitable for exploratory search or when a user wants to study a particular topic in detail. It is frequently employed by users of text archives – having references or links to other related web pages, articles and books. The linking concept became more popular with World Wide Web where it become a major navigation approach (Berners-Lee and Cailliau, 1990) and its popularity in multimedia is still growing (Tan et al., 2011).

Hyperlink is defined as "an electronic link providing direct access from one distinctively marked place in a hypertext or hypermedia document to another in the same or a different document" (Merriam-Webster Online, 2009). The source marked place of the link (e.g. relevant text, relevant part of an image) is called the *anchor*. This work focuses on hyperlinks in video archives. Both the anchors and targets are video segments defined by the source video recording and its

beginning and ending playback times (Ordelman et al., 2015a). An example of hyperlinking in a video archives is depicted in Figure 4.1.

Figure 4.1: Example of Hyperlinking in the BBC 2014 SH set. The anchor segment and retrieved segments are all represented by single keyframes.

In contrast to common hyperlinking settings where anchors and targets are defined by the author of the archive, an anchor can be any segment of the video selected by the user for the purposes of this thesis. Unlike with text, anchors in multimedia are still rare. For example, they appear in YouTube videos which contain additional annotations of the videos displayed at particular times defined by the video creators. Anchors also appear in the audio distribution platform SoundCloud[1], which allow listeners to comment at any particular time of the audio track.

Recommender systems

Hyperlinking is closely related to recommender systems (Bhatt et al., 2014), which typically provide suggestions for items to be used as anchors (Ricci et al., 2010). Hyperlinks can also be considered as recommendations of video segments to watch. Ordelman et al. (2015a) state that the main difference between video hyperlinking and other related techniques such as video recommendation (Ricci et al., 2010) or near-duplicate detection (Furht, 2008) is that video hyperlinking focuses on "'give me more information about this anchor' instead of 'give me more based on this anchor or entity'". Thus the attention is focused on semantic similarity between the anchoring and target segments. This is typical for texts

[1] `https://soundcloud.com`

where links usually point to topically related or explanatory content. Since a number of modalities exists in videos, hyperlinking in audio-visual content is a comparatively more complex problem. Segments may describe the same topic, visualise the same topic, the topic described in the anchoring segment may be visualised in the target segment, the same composition may be played in both segments (or even the same artists may play different compositions), the same person, sound or noise may appear in both segments.

In opposition to hyperlinking, which is primarily intended for professional media content producers, video recommendation systems are mainly intended to entertain (Davidson et al., 2010) and increase the number of video views (Gomez-Uribe and Hunt, 2015). Therefore, recommender systems typically do not analyze video content but they widely use metadata, collaborative filtering approaches (Youtube) (Baluja et al., 2008; Bendersky et al., 2014) and manual recommendations (TED) (Pappas and Popescu-Belis, 2013). However, access to the information stored in videos using recommendations can be limited. Metadata, such as title and video description, must be generated manually and may be unreliable. Even if the metadata is carefully crafted, e.g. in TV program archives, it can hardly describe the full content of the video. User statistics needed for collaborative filtering are also often unavailable, especially in the case of a new user's cold start (Schein et al., 2002) or when a new video needs to be processed.

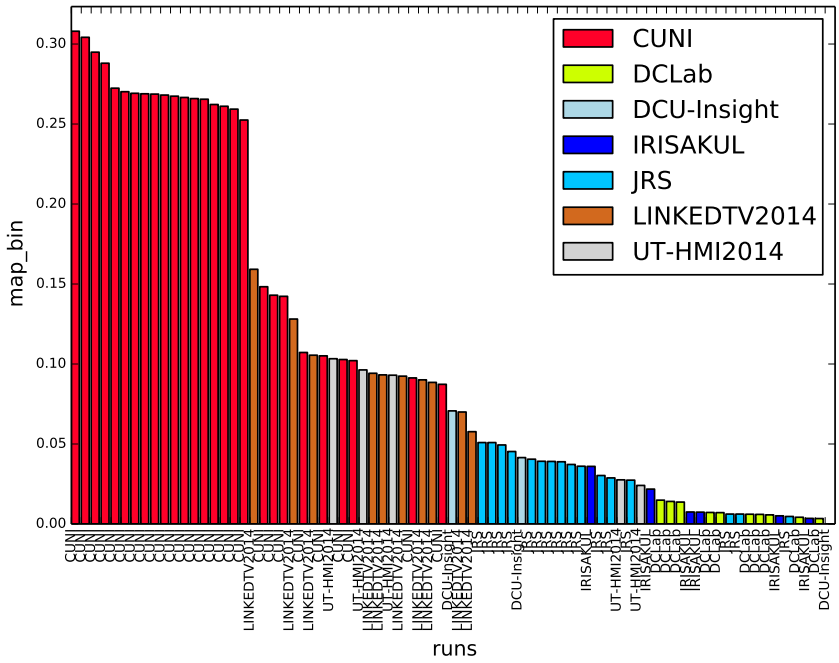
Chapter overview

In the first section, hyperlinking approaches are described in general. Utilization of audio and visual modalities are then described in following sections, respectively. Effectuated experiments were performed in the 2014 SH Task, SSSS Tasks and 2015 Video Hyperlinking Task and use BBC 2014 SH set and SSSS set. The experiments performed best in the comparison with other teams participating in the Hyperlinking sub-task of the 2014 SH Task, see Figure 4.2[2] (Eskevich et al., 2014a).

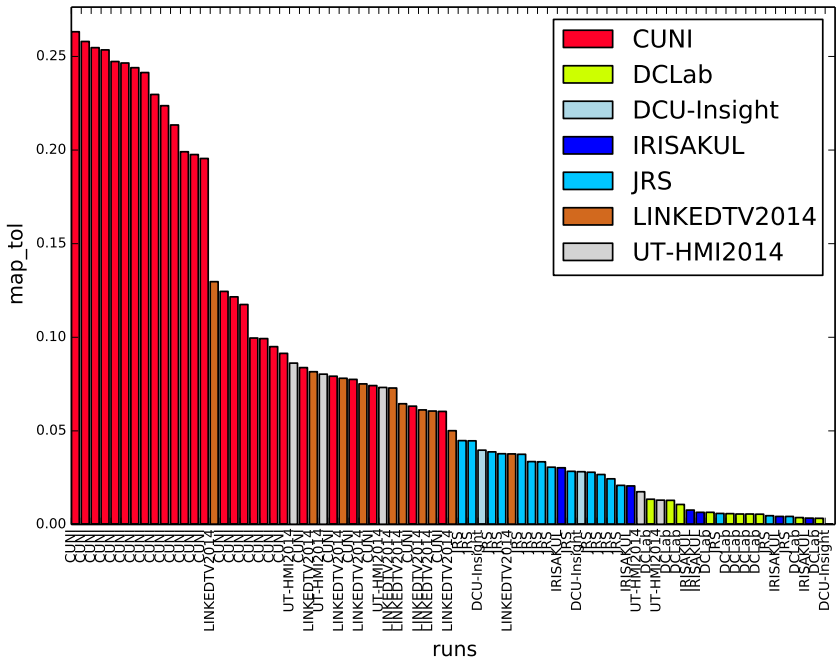
4.1 Hyperlinking Experiments

This section describes the hyperlinking approach in general. Data, evaluation measures and methods used in different evaluation campaigns are described here in detail.

[2] Figures were provided by the task organizers.



(a) MAP-bin measure



(b) MAP-tol measure

Figure 4.2: A comparison of the team results from the Hyperlinking sub-task of the 2014 SH Task. Our team results are labeled as *CUNI*.

Hyperlinking benchmarks

Hyperlinking has been addressed at several evaluation campaigns. A related problem which explored linking of video segments to relevant Wikipedia[3] pages was solved at the *Finding Related Resources Across Languages* task at the *VideoCLEF 2009* Benchmarking campaign. The task involved a multilingual aspect as the source anchors created from Dutch videos were linked with English Wikipedia pages. The *Beeldenstorm* dataset consisting of art documentaries was used in this task.

The Hyperlinking task, as described in the previous sections, was explored in the SH Task organized at the *MediaEval* Benchmark between 2012 and 2014. Since 2015, it has been a part of the *TRECVID* as the *Video Hyperlinking* Task. Datasets used in these tracks were described in Section 2.1. The Blip.tv dataset was used at MediaEval 2012. The ground truth segments from the test set used in the Search sub-task were used as the anchor query segments for hyperlinking. Additionally, task participants were allowed to define their own anchoring segments by using the segments retrieved by their text-based search systems. This simulated the scenario in which a user seeks additional details about segments retrieved in the text-based search. Blip.tv videos were again used in the 2016 TRECVID Video Hyperlinking task. However, this task was more focused on multimodality (e.g. speaking about visual content) and uploader intent (e.g. convey knowledge, teach practice, illustration).

The BBC 2013 SH set and the BBC 2014 SH set were used in 2013 and 2014 SH Task, respectively. Anchors were created by participants of the user study at BBC. In 2013, the user study participants were asked to find short segments which were somehow interesting to them. These segments were then used in the Known-Item Search task as the target segments. Moreover, participants were also asked to mark hyperlinking anchors within these known-item segments. The target segments clearly could not overlap with the anchors and their lengths were restricted to between 10 and 120 seconds.

In 2014, the focus was on a home user scenario. Participants were asked to define anchors using iPads. They first browsed the dataset to become familiar with it. Participants then searched the dataset using a text-based search, watched retrieved segments and marked anchors within these retrieved segments. Participants were also asked to describe the types of links they would like to see based on marked anchoring segments. This information was available during the evaluation but was hidden from the participants. Retrieved segments overlapping with query segments were eliminated and only anchors longer than 5 seconds, clearly

[3] `https://www.wikipedia.org`

Task	Training anchors	Test anchors
SSSS 2013	198	29
SH 2014	50	30
Video Hyperlinking 2015	30	100

Table 4.1: Statistics of the anchors used in different tasks

satisfying user expectations, and for which similar segments possibly exist in the dataset, were used in the task (Eskevich et al., 2014a).

The BBC 2014 SH set was again used in the 2015 TRECVID Video Hyperlinking Task. However, this task was directed primarily at media professionals and content creators. The anchors were created by journalism students and media professionals who were asked to identify segments which they would like to enrich with hyperlinks if they were content producers (Over et al., 2015). Unlike previous years, links within the same video file were not allowed. Statistics of the number of anchors used in tasks organized between 2013 and 2015 are displayed in Table 4.1[4]

Similar Segments in Social Speech

The 2013 MediaEval *Similar Segments in Social Speech* Task focused on problems related to hyperlinking. The task scenario considers a hypothetical user trying to locate information in video archive, e.g. a new university student seeks information about a particular course from a dataset of recorded interviews between students. When the user finds some relevant information, he or she then wishes to expand the search to locate additional relevant information. The dataset provided in the task contained manually detected segments grouped into similarity sets based on the segments' topic (e.g. TV-shows, travel, planning-class-schedules). According to the hyperlinking scenario, each segment marked in the dataset can thus form the anchor and all segments within the similarity set with the anchor become target segments.

Hyperlinking Evaluation

Predefinition of the reference set to be used for evaluation is a complex task, especially for large datasets. Thus, hyperlinking is typically evaluated manually using crowdsourcing. Survey participants are asked to watch the anchor and retrieved segments and state whether the retrieved segments describe the anchor in detail. Evaluation measures are typically identical to those used for the

[4] Only tasks with results stated in this chapter are displayed.

evaluation of text-based retrieval described in Section 2.2. Specifically, MAiSP, MAP-bin and MAP-tol measures have been recently used.

Since the number of retrieved segments is very large, only the top retrieved segments (e.g. top 10 segments) for each run are pooled and evaluated. Therefore, experiments described in this chapter, which were performed after an official evaluation, were evaluated with respect to the set of officially evaluated segments. The reference set thus only contains results which were officially submitted by any participating team and ranked among top results. If the additional runs do not overlap with any relevant segment from the reference set, they cannot be considered to be relevant. Runs officially submitted to the task can thus be considered to be superior compared to the additional runs.

4.1.1 From text-based search to video hyperlinking

In the presented experiments, textual similarity based on subtitles and automatic transcripts described in the previous section are combined, with similarity being based on visual and acoustic features. This approach 1) analyzes the content of the recordings, 2) enables users to retrieve relevant segments from the video recordings, and 3) enables the combination of different modalities (metadata, speech, visual and acoustic content).

In the described approach, the anchor is converted into a textual query. All words in the subtitles or transcripts which are between boundaries of the anchor are used to form the textual query. Since a relevant video segment is requested in the output, recordings are, as in the text-based search, presegmented into shorter passages which can be again converted to text. The anchor can then be used in IR to acquire the most similar segments and the hyperlinking is thus converted into text-based search. Acoustic and visual information are then either appended as textual information to each segment (*early fusion*) or the score achieved by the text-based search is combined with the score of the similarity between the anchor and output segment based on the acoustic or video information (*late fusion*).

4.1.2 Other hyperlinking approaches

A full range of methods have been applied to hyperlinking. Presegmentation of videos is the dominant approach to find relevant segments. Researchers use fixed length segments measured by time (Cheng et al., 2015; Niaz et al., 2015; Chen et al., 2013) or the number of included words (Levow, 2013), sentence boundaries (Chen et al., 2015), shots (Lokaj et al., 2013; Paróczi et al., 2014), visual similarity between frames (Sahuguet et al., 2013) and boundaries defined by the TextTiling algorithm (Pang and Ngo, 2015; Bhatt et al., 2014). Nies et al.

(2013) optimize segment length in order to maximize similarity between query and target segments. Other researchers use hierarchical topic segmentation, where the segment boundaries are recalculated using lexical cohesion and disruption (Simon et al., 2014). Similarly, Le et al. (2014) gradually group adjacent segments which are semantically similar either in terms of visual similarity or lexical cohesion.

However, some researches localize the jump-in points directly without any presegmentation. For example García et al. (2013) placed jump-in points at the location that maximized their similarity function based on shared words and WordNet distances between words. Similarly, Preston et al. (2013) represent each video by a probability density function and the points with the highest probabilities become the jump-in points.

Methods also differ in types of modalities which are used in the retrieval. Some researchers only use textual information from transcripts and subtitles (Levow, 2013; García et al., 2013). Textual information can also be enriched by synonyms (Paróczi et al., 2014) or named entities (Nies et al., 2013; Chen et al., 2013). Textual query expanded by connected entities from Freebase[5], geographical entities from GeoNames[6], synonyms from WordNet, and DBpedia[7] words were used by Lokaj et al. (2013). Word vectors (Mikolov et al., 2013) can also be used as word representations. Pang and Ngo (2015) prove that this approach can be helpful when word2vec is combined with TF IDF. Manually crafted video metadata can be added to subtitles and transcripts. Chen et al. (2014) use Linear Discriminant Analysis to tune weights for combinations of text-based retrieval and retrieval based on video metadata. In their following work (Chen et al., 2015), they combine these retrieval outputs by fusion with weights acquired using the Maximum Deviation Method (Wilkins, 2009).

Textual and visual information are often combined. Recent research efforts mostly employ visual concepts (Sahuguet et al., 2013; Xu et al., 2015; Le et al., 2014). Simon et al. (2015b) use bi-lingual LDA to create a probabilistic translation model between lexical transcripts and visual concepts. The results acquired by text-based retrieval using subtitles and transcripts can also be re-ranked using SIFT descriptors (Lokaj et al., 2013). Niaz et al. (2015) employ lexical context of visual features using the word2vec algorithm. On the other hand, the approach used in the SSSS Task by Werner and Ward (2013) is purely based on prosodic information (76 local prosodic features over 6-second long sliding window). Using prosodic information may benefit searches for similar dialogue activity such as questions, telling stories, surprising statements and agreement.

[5] `https://developers.google.com/freebase`

[6] `http://www.geonames.org`

[7] `http://wiki.dbpedia.org`

However, the extent of possible improvements achieved using multiple modalities is not clear. Even though, some researchers report an improvement using a combination of modalities, the best scoring approach at the 2015 Video Hyperlinking Task (Cheng et al., 2015) used only textual data. Researchers state that multimodal features do not provide any improvement although a broad range of visual and audio features were used in their experiments.

4.1.3 Baseline system

In the presented experiments, Passage Retrieval using the Terrier Framework with its associated Hiemstra LM implementation was applied to the recordings to facilitate segment retrieval. Documents were divided into 50-second long segments with new segments being created every 10 seconds. Text lying within the segment boundaries was then indexed. Similarly, a textual query was created from a query segment by utilizing all words in the subtitles lying within the query segment. Text-based retrieval was then applied to the dataset.

For this work, the Terrier parameter was set to 0.35 and the Porter stemmer and Terrier's stopwords list were also used. In all experiments, retrieved segments were post-filtered and segments which partially overlapped with either the query segment or another higher ranked retrieved segment were removed ("overlap removal" filtering).

4.2 Audio information

A detailed discussion of speech transcript utilization is presented in this section. Specifically, the three most significant problems associated with automatic transcripts are addressed: 1) restricted vocabulary used in the ASR system 2) lack of transcripts' reliability and 3) lack of content (Larson and Jones, 2012b). Performed experiments are schematically depicted in Figure 4.3.

Restricted ASR vocabulary results in words being omitted from the transcripts or being misrecognized. The number of unique words in automatic transcripts is almost three times smaller than the number of unique words in subtitles in the presented BBC 2014 SH set. Moreover, the low frequency words are expected to be the most informative for IR purposes. This out-of-vocabulary word problem may be partially overcome by expansion of the data segments and query segments or by combining different transcripts. In the presented experiments, data and query segments are expanded by metadata. Metadata possibly contain names and designations which are important for a particular recording but which

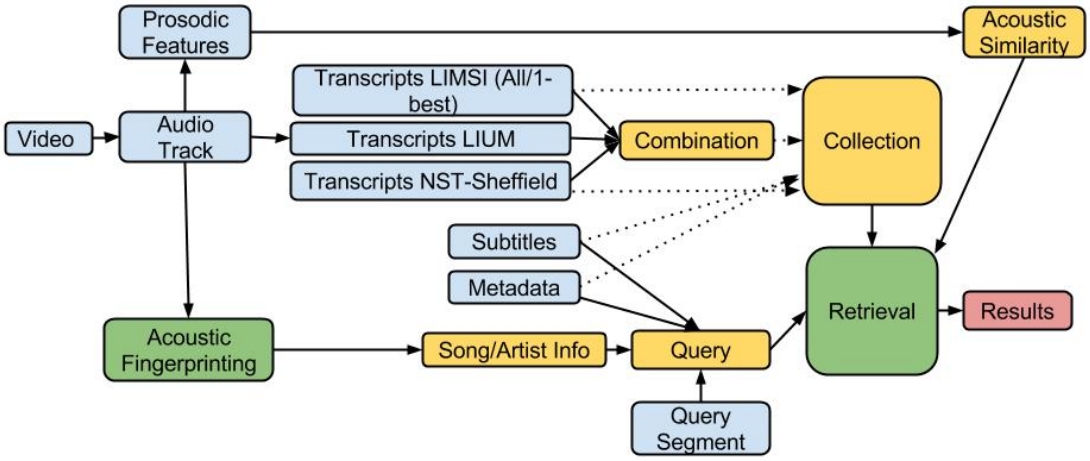


Figure 4.3: Scheme of the performed experiments with audio information. Blue items are contained in the input.

rarely occur in the general vocabulary. Anchor segments are also expanded by the content surrounding the query. Additionally, three ASR systems are analyzed and compared to human subtitles.

Even though the quality of the automatic transcripts has improved rapidly over the last few years, **transcript reliability** still needs to be addressed. The reliability is especially a problem if the quality of the recording is lower or if a number of available language resources for particular language is lower. To address transcript reliability, utilization of the first, most reliable word and all its word variants are compared. In addition, only words with high confidence scores are used.

Even if speech information is accommodated correctly, information from other modalities is still missing in the retrieval and the retrieval thus suffers from a **lack of different content**. This problem is partially addressed by employment of "*acoustic fingerprinting*" and "*acoustic similarity*". Acoustic fingerprinting obtains additional information from the music contained within the query segment. It may be especially helpful for hyperlinking music programmes which are likely to occur in TV archives (e.g. artist's music clips can be linked to artist's interview). "Acoustic similarity" further extends retrieval by acoustic features.

Experiments presented in this section are performed using BBC 2014 SH set. Some of the presented experiments were submitted to the BBC 2014 SH Task and evaluated officially by the task organizers. Specifically the experiments without any data and query expansion (row 1 in Table 4.3 for each transcript),

the experiments with data and query expanded (row 4 in Table 4.3 for each transcript) and the experiments for the acoustic similarity (Section 4.2.5) were submitted to the task and officially evaluated. The rest of the experiments was performed after the official evaluation.

4.2.1 Transcript quality

The BBC 2014 SH set contains three automatic transcripts provided by LIMSI, LIUM and NST-Sheffield (see Section 2.1). Lamel (2012) reports Word Error Rate (WER) of the LIMSI ASR system obtained on two datasets consisting of a mixture of English Broadcast News and Broadcast Conversation as 17.3% and 20.1%. The lowest WER achieved by the LIUM transcripts on the test set consisting of TED Talks was 11.1% (Rousseau et al., 2014). The NST system, designed for transcribing the multi-genre BBC archive (Lanchantin et al., 2013), was evaluated on BBC data consisting of radio and television drama programmes. Its lowest achieved WER was 27.1%.

Stated WERs are calculated for different datasets and thus could not be directly compared. To compare ASR quality on BBC 2014 SH set used in the experiments, 10 documents were randomly selected[8] from the test data and WER was calculated for these documents. Transcripts were unified (in terms of punctuation and letter case) and additional functional and special words were removed (e.g. *{fw}*) and compared to subtitles[9]. However, the transcripts may still differ in spelling (e.g. *ok* vs. *okay*, and *3* vs. *three*). The comparison of calculated WERs appears in Table 4.2. The most common errors are similar for all transcripts (confusions between *a* and *the* and between *it's* and *is*).

Word variants

The LIUM transcripts consist of single best word variants with each word having an assigned confidence score. Similarly, only the single best words are given in the NST transcripts but no confidence scores are given. In contrast, LIMSI transcripts contain several word variants occurring at the same time and similar to the LIUM transcripts, the confidence of each word variant is given.

Although the subtitles are reliable and their vocabulary is not restricted, timing information is only available for a block of text appearing at one time on the screen. Therefore, the time of the utterance of a particular word must be

[8] The WER could not be calculated on full collection due to technical restrictions.

[9] The subtitles may contain additional words such as captions and production information at the end of the program, which may increase WER, but this disadvantage is the same for all examined transcripts.

Transcripts	Words Train		Words Test		WER(%)
	Unique	Total	Unique	Total	
LIMSI All	71k	13M	84k	27M	71.8
LIMSI 1-best	66k	11M	79k	22M	57.5
LIUM	95k	9M	122k	18M	65.1
NST	55k	9M	59k	17M	58.6
Subtitles	227k	13M	313k	24M	–

Table 4.2: Word counts of the BBC 2014 SH train and test sets and WER of automatic transcripts calculated on 10 programmes randomly selected from the BBC 2014 SH set.

approximated. Unlike subtitles, the exact time stamp of the utterance of each word is given in the automatic transcripts.

4.2.2 Query Expansion

The BBC 2014 SH set contains manually-created metadata such as the title and a description of the program. These data (specifically title, episode title, description, short episode synopsis, service name, and program variant) is used to concatenate each data segment with metadata of the corresponding file. In detail, each word occurrence from the data segment and each word occurrence from metadata were used to create a new segment. Similarly, each query segment was concatenated with corresponding metadata. Additionally, the context of the query segment was expanded using 200 seconds before and after the query segment. The context was concatenated with the query segment in the same way as metadata. The length of the context was tuned on the development data for the highest MAP measure.

The results for various transcripts, expansion of the data and query segments by metadata and context are shown in Table 4.3. Expanding the queries and segments with keywords from metadata and context significantly increased the scores achieved by subtitles and all automatic transcripts. This is in line with the results achieved for metadata utilization in text-based search in Section 2.3.2.

The greatest improvement was achieved when both query and data segments were concatenated with metadata and query segments were extended by the context. In this case, programmes from the same episode or programmes broadcast on the same day (the date is in some cases part of the title) can be promoted. When no metadata is utilized, the performance corresponds with the WER relatively well, and the subtitles clearly outperform the automatic transcripts. However, the metadata and context produce a much higher relative improvement to

Transcript	Data Expansion	Query Expansion	MAP-bin	MAP-tol
Subtitles	None	None	0.142	0.122
Subtitles	Metadata	None	0.152	0.127
Subtitles	None	Metadata+Context	0.258	0.218
Subtitles	Metadata	Metadata+Context	**0.269**	**0.247**
LIMSI All	None	None	0.103	0.074
LIMSI All	Metadata	None	0.101	0.080
LIMSI All	None	Metadata+Context	0.196	0.154
LIMSI All	Metadata	Metadata+Context	**0.266**	**0.230**
LIUM	None	None	0.087	0.060
LIUM	Metadata	None	0.094	0.068
LIUM	None	Metadata+Context	0.173	0.146
LIUM	Metadata	Metadata+Context	**0.260**	**0.255**
NST	None	None	0.102	0.080
NST	Metadata	None	0.106	0.084
NST	None	Metadata+Context	0.197	0.158
NST	Metadata	Metadata+Context	**0.261**	**0.224**

Table 4.3: Results for different transcripts, for employment of metadata and context of the query. The best results for each transcript are in **bold type**.

the automatic transcripts than to the subtitles. The improvement to the subtitles is 90% and 103% for MAP-bin and MAP-tol measures, respectively, and 197% and 322% on the LIUM transcripts. When metadata and context are utilized, to the LIUM transcripts even outperform the subtitles in MAP-tol measure; the overall highest MAP-tol score is achieved in this case. This result confirms that expansion using metadata and context can substantially reduce the effects of a reduced vocabulary, even though the transcripts are trained on different data types and have a relatively high WER.

4.2.3 Combination of transcripts

Another approach which helps in dealing with the problem of reduced vocabulary is to combine different transcripts. One possible approach to transcript combination is to join them into a single document. All words in selected transcripts which lie within particular segment boundaries are then merged[10]. These experiments are tabulated in Table 4.4.

[10] All words in all merged segments are indexed.

Transcript	MAP-bin	MAP-tol
Subtitles	0.269	0.247
LIMSI	0.266	0.230
LIUM	0.260	**0.255**
NST	0.261	0.224
LIMSI + LIUM	0.271	0.242
LIMSI + NST	**0.275**	0.231
LIUM + NST	0.269	0.233
LIMSI + LIUM + NST	0.274	0.230

Table 4.4: Results for combination of transcripts. Metadata and context are utilized and all word variants from the LIMSI transcripts are used. Best results are in **bold type**.

The overall highest MAP-bin score is achieved using the union of the LIMSI and NST transcripts; the combination even outperforms the results achieved with subtitles. Combination of the transcripts is generally helpful (e.g. the combination of the LIUM and NST transcripts is in terms of the MAP-bin score significantly better than both LIUM and NST transcripts). But as the LIUM transcripts already perform well and achieve a high MAP-tol score, combination with other transcripts adds confusion, lowers the transcript quality and the score drops.

4.2.4 Transcript reliability

Reliability of the transcripts is improved by using different **word variants** created by the ASR system. Word variants are available for selected words in the LIMSI transcripts; all of these variants have an assigned confidence score. Employment of all available words was thus compared to the case when only the most confident word was used. The MAP-bin score slightly rose from 0.266 when all available words were used to 0.268 when only the most confident word was used. The MAP-tol score rose from 0.230 to 0.232 in the same case.

Only words from LIMSI and LIUM transcripts with a **confidence score** higher than a given threshold were also used in order to improve transcript reliability. The threshold confidence scores for both transcripts were tuned with the train set. The threshold on the LIUM transcripts was set to 0.7 and 0.8 for MAP-bin and MAP-tol measures, respectively, and to 0.8 and 0.6 for MAP-bin and MAP-tol scores respectively on the LIMSI transcripts. Although this strategy increased both scores on the train set, it did not outperform the subtitles on the test set. In addition to filtering the results based on confidence, voting was also

employed. Each word was only indexed if it exists in at least two transcripts out of three. Also according to Kittler et al. (1998), voting should guarantee higher reliability of the indexed words. However, this approach also did not prove to be helpful with the presented data.

4.2.5 Acoustic similarity

The audio track contains additional information which can help identify the content or genre of the recording. Acoustic similarity between short audio tracks may not only help to retrieve segments containing audio tracks identical to those in the query segment (e.g. signature tunes and jingles) but can also be useful for detection of semantically related segments (e.g. segments containing action scenes and music). Audio information can also be helpful in recognizing the mood and emotions of a speaker.

Eight prosodic features were provided in the dataset (energy, loudness, voice probability, pitch, pitch direction, direction score, voice quality, and harmonics-to-noise ratio) (Eskevich et al., 2014b). Acoustic similarity between data segments and query segments was calculated based on these features and these results were then combined with textual similarity between segments calculated on lexical information acquired from subtitles.

The prosodic features were calculated for every 10ms of data playback time. The most similar sequences of vectors near the beginning of each data segment were found for each overlapping sequence of 10 prosodic 8-dimensional feature vectors appearing up to 1 second from the beginning of each query segment. The strategy is displayed in Figure 4.4.

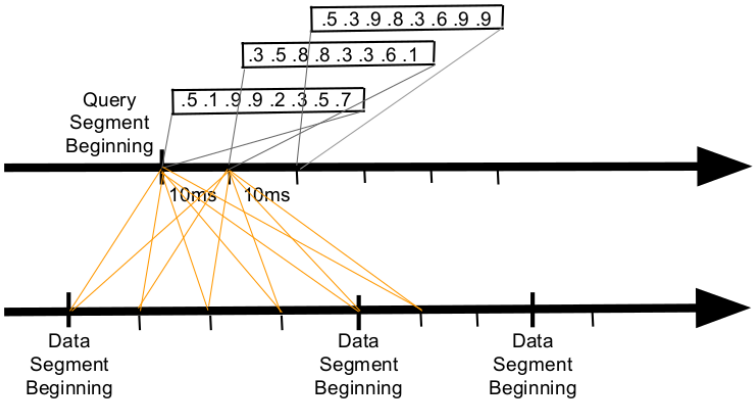


Figure 4.4: Calculation of the acoustic similarity between query and data segments.

Similarity between data and query vector sequences was calculated as the sum of differences between the corresponding vectors of the sequence. These vector differences were calculated as the sum of the absolute values of the differences between the corresponding items of the prosodic vectors (i.e. the L1 distance). The difference of each item of the prosodic vector was normalized to have component values between 0 and 1 to ensure that all prosodic features had equal weights.

The highest acoustic similarity between segments was linearly combined with the text-based similarity score calculated on the subtitles:

$$\begin{aligned}\mathit{FinalScore}(\mathit{segment/query}) = \mathit{TextSimilarity}(\mathit{segment/query}) \\ + \mathit{Weight} * \mathit{AcousticSimilarity}(\mathit{segment/query}).\end{aligned} \quad (4.1)$$

The *Weight* was tuned on the train data.

Acoustic Similarity decreased the MAP-bin score very slightly from 0.2689 to 0.2687, but the MAP-tol score rose from 0.2465 to 0.2473. Therefore, application of audio similarity looks promising, but it needs to be further explored since despite the computational restrictions, the MAP-tol measure increased.

4.2.6 Acoustic fingerprinting

Shazam[11] and SoundHound[12] are well known services which enable users to identify recorded music compositions. Since the BBC 2014 SH set contains musical selections either as the main object in a query segment or in the background, the identification of an artist and composition title could provide additional beneficial information about the query segment which may be used for further expansion of the query. Doreso[13] music fingerprinting service was utilized to provide such identification information, since it offers developers' API. We also experimented with EchoNest[14] service but it retrieved no supplemental information for the train queries.

Since the Doreso service employs fingerprinting techniques and works well for near duplicates, it is essential that the submitted passages contain as few interfering sounds as possible, such as speech. The query segments were divided into 10-second long passages; new passages were created each second. These created passages were submitted to the service. Doreso retrieved the composition title and artist name for 4 queries out of 30 from the train set and 10 queries out of 30 from the test set. Query segments were then concatenated with the

[11] `http://www.shazam.com`

[12] `http://www.soundhound.com`

[13] `http://developer.doreso.com`

[14] `http://the.echonest.com`

corresponding composition title and artist and album names, when available, but doing so caused both retrieval scores to drop.

The fact that music identification information did not increase the scores may be partially due to evaluation methods which do not consider this aspect of the retrieval. Even though music identification may add interesting information, concatenation of the query segment with composition and artists' names may be insufficient. Utilization of additional identification of data segment musical compositions would probably be beneficial, but is not easily realizable due to the size of the dataset. Another problem with music identification is that it only works well with near duplicates. – Doreso works well on background music, but it cannot be expected to perform well in the case of live performance versions of the composition.

4.2.7 Conclusion

This section overviews our experiments with various audio content of audio-visual recordings. We compare subtitles and three automatic transcripts with different quality. We conclude that the transcript quality influences the quality of the retrieval but it is possible to completely deal with this disadvantage using additional content such as metadata and context of the segment. Disadvantage of restricted vocabulary of the transcripts can also be addressed by combining different transcripts. Using transcribed word variants, if they are available, brings additional noise to the retrieval and using single best transcription thus improves reliability. Employing confidence scores of the transcribed words and voting of the transcripts bring no improvement. Combination of the transcripts and acoustic similarity, calculated using prosodic features, and recognized music compositions did not improve the calculated scores. However, this kind of information can be still very helpful, especially if users are interested in music or sound in the query segment.

4.3 Visual information

This section surveys various state-of-the-art visual processing methods and describes their utilization in hyperlinking. Visual information calculated using Feature Signatures (Rubner and Tomasi, 2001), SIMILE descriptors (Kumar et al., 2009) and Convolutional Neural Networks (CNN) (Krizhevsky et al., 2012), is used to quantify similarity between video frames and to find similar faces, objects and settings. Visual concepts in frames are also automatically recognized and textual output of the recognition is combined with search results based on

subtitles and transcripts. All presented experiments were performed in the 2014 SH Task and 2015 Video Hyperlinking Task.

First, methods for calculating similarity between frames tested in 2014 SH Task are described. Retrieval which utilizes visual similarity employing Feature Signatures, which acquires information on the frame setting, is compared with retrieval methods based purely on subtitles and transcripts. Results are analyzed using the BBC 2014 SH set. In further experiments, Feature Signatures are compared with deep-learning based methods and face similarity algorithms are utilized. Finally, the most promising approaches are compared in the 2015 Video Hyperlinking Task.

4.3.1 Visual Descriptors in the 2014 SH Task

An overview of various visual descriptors (Feature Signatures, CNN and face descriptors) and methods of combining them with text-based retrieval (linear combination of weights and query expansion in which words from a query are used along with detected concepts) is displayed in Figure 4.5. Experiments using these descriptors were performed in the 2014 SH Task and thereafter on the same dataset. Experiments reported in Table 4.5 were officially evaluated in the task, the rest of the experiments reported in this section was evaluated afterwards.

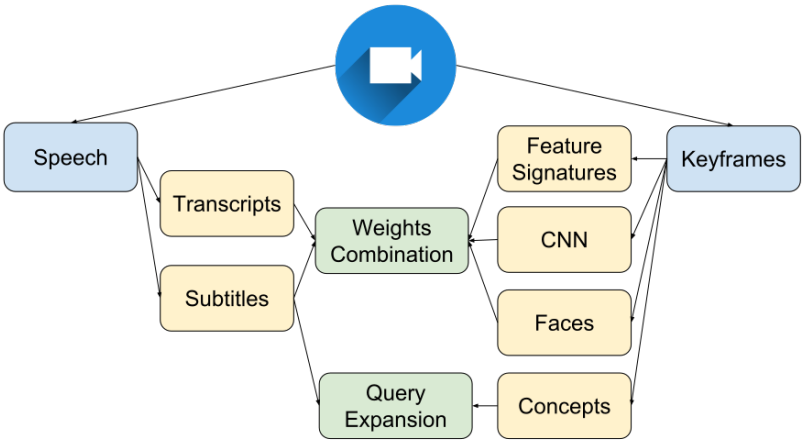


Figure 4.5: Visual methods overview.

Feature Signatures

In order to represent video frames, *Feature Signatures* that approximate distribution of color and texture in the image can be used. Unlike modern CNN descriptors, excellent in recognizing specific objects, this traditional descriptor

can be used to identify keyframes having a similar background and setting. Feature Signatures have recently been used in known-item search tasks (Cobârzan et al., 2017). The authors have created an efficient and effective sketch-based retrieval system (Blažek et al., 2014; Blažek et al., 2015), where each keyframe is represented by Feature Signatures and users can query the database of Feature Signatures using simple signature-based sketches.

Visual similarity

Visual similarity between the query segment and each data segment was calculated using Feature Signatures. A distance was calculated between each keyframe in the query segment and each keyframe in the data segments. This process is displayed in Figure 4.6. Automatically detected keyframes representing the shots provided by the task organizers were used in the presented experiments. A distance between segments was then calculated as a minimal similarity $(1 - \mathit{distance})$ between keyframes in the query and data segments. Finally, the visual similarity of the segments was linearly combined with their textual similarity score acquired from the text-based retrieval:

$$\begin{aligned}\mathit{FinalScore}(\mathit{segment/query}) = \mathit{Score}(\mathit{segment/query}) \\ + \mathit{Weight} * \mathit{visualsimilarity}(\mathit{segment/query}),\end{aligned} \quad (4.2)$$

and the retrieved segments were scored according to the Final Score. The Weight was tuned on the train data and set to favor results of text-based retrieval.

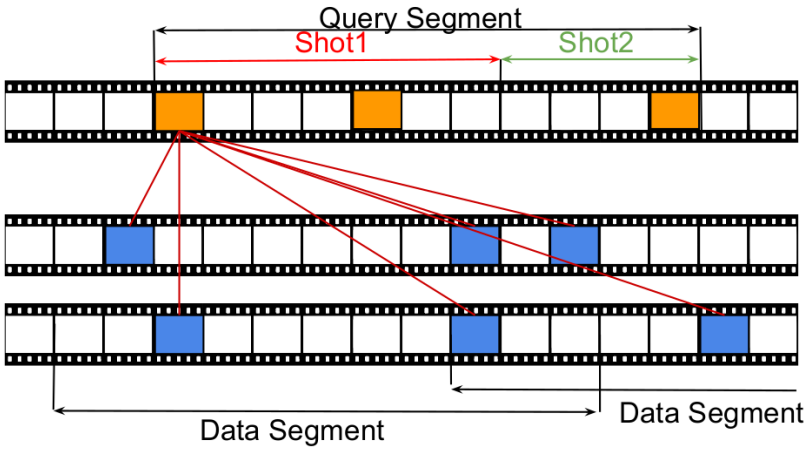


Figure 4.6: Calculation of the visual similarity between segments.

Feature Signatures definition

Position-color-texture Feature Signatures (Rubner and Tomasi, 2001; Kruliš et al., 2012; Kruliš et al., 2016) were utilized to approximate a distribution of color and texture in each keyframe. This descriptor can be utilized in image retrieval tasks, where color and texture are relevant. Formally, given a feature space $\mathbb{F}$, the *Feature Signature* S^o of a multimedia object o is defined as a set of tuples $\{\langle r_i^o, w_i^o \rangle\}_{i=1}^n$ from $\mathbb{F} \times \mathbb{R}^+$, consisting of representatives $r_i^o \in \mathbb{F}$ and weights $w_i^o \in \mathbb{R}^+$. Each keyframe is thus represented by a set of Feature Signatures which describe color of image regions. An example of an image represented by Feature Signatures is displayed in Figure 4.7.

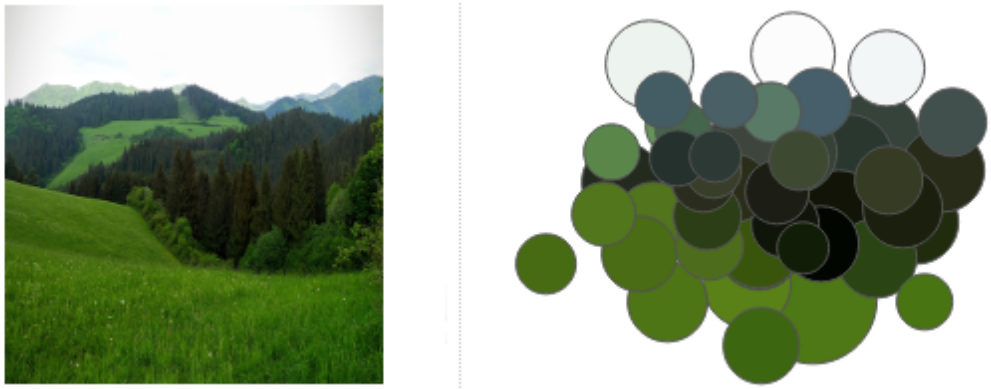


Figure 4.7: Representation of an image using Feature Signatures.

The distance between Features Signatures is calculated using the Signature Quadratic Form Distance. Since the Signature Quadratic Form Distance is a Ptolemaic metric, metric / Ptolemaic indexing techniques can be utilized for efficient retrieval (Hetland et al., 2013; Beecks et al., 2011). Furthermore, GPU implementations already exist, enabling both efficient extraction of feature signatures and evaluation of the Signature Quadratic Form Distance (Kruliš et al., 2012, 2013).

Employing Feature Signatures enables visual similarity to work exceptionally well in detecting similar settings and backgrounds. This is particularly important in working with TV archives, in which a similar background occurs throughout a series. However, Feature Signatures can fail to detect some details in keyframes, e.g. it is not possible to recognize a particular person. Feature Signatures-based distances were provided by the Department of Software Engineering at Charles University.

2014 SH Task results

The hyperlinking text-based retrieval system described in Section 4.1.3 with metadata and context expansion was used as the baseline. Comparison of results using visual similarity employing Feature Signatures and text-based retrieval is displayed in Table 4.5.

Transcript	Weights	Filtering	MAP-bin	MAP-tol
Subtitles	None	ROO	0.269	0.247
Subtitles	Visual	None	**0.308**	0.100
Subtitles	Visual	ROO	0.272	**0.258**
LIMSI	None	ROO	0.266	0.230
LIMSI	Visual	None	**0.304**	0.100
LIMSI	Visual	ROO	0.269	**0.241**
LIUM	None	ROO	0.259	0.255
LIUM	Visual	None	**0.288**	0.099
LIUM	Visual	ROO	0.262	**0.263**
NST-Sheffield	None	ROO	0.261	0.224
NST-Sheffield	Visual	None	**0.295**	0.091
NST-Sheffield	Visual	ROO	0.266	**0.244**

Table 4.5: Results with and without use of visual similarity for different transcripts with "overlap removal" filtering. Best results for each transcript are in **bold type**.

The best results for both measures were achieved when visual similarity was employed. The highest MAP-bin score was achieved on the subtitles when segments overlap. Surprisingly, the highest MAP-tol score was achieved on the LIMSI transcripts when the segments did not overlap. This may be caused by the precise timing given for each word in the LIUM transcript. The timing of each word in the subtitles was approximated based on the utterance's beginning and ending times. In the case of the 15-second long tolerance windows, this approximation could decrease the score. The MAP-bin score is generally higher when the segments can overlap; the MAP-tol scores were higher when overlapping segments were filtered out. Low MAP-tol scores achieved when overlapping segments were not filtered out were caused by the fact that the MAP-tol measure takes into account each relevant segment only once. Overlapping retrieved segments thus lowered the score.

Visual similarity proved to be helpful in the task. Even though the improvement in terms of MAP-bin score was small, Feature Signatures improved the results consistently for all transcripts in both measures. The presented system

achieved the highest results in the Hyperlinking sub-task of the 2014 SH Task (Eskevich et al., 2014a).

Visual similarity analysis

Three examples in which Feature Signatures improved the retrieval quality are shown in Figure 4.8.

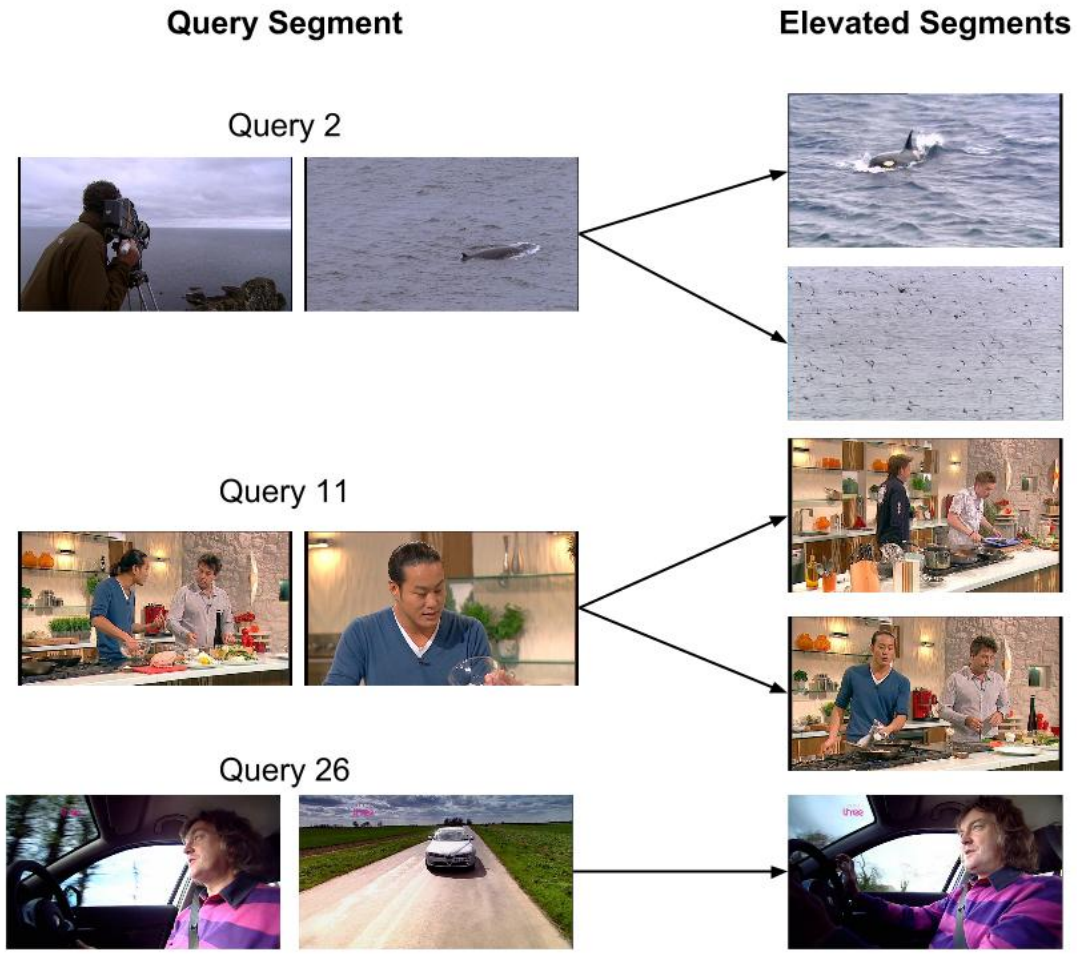


Figure 4.8: Examples in which employment of visual similarity increased the retrieval quality.

Color-position-based Feature Signatures work well for types of queries similar to Query 2. Even though query and target keyframes are not necessarily from the same TV series or program, they are similar in terms of color distribution throughout the image. Similarly, the results in Query 11 are improved due to the retrieval of the same setting. In Query 26, the only top relevant retrieved segment was the segment before the beginning of the query segment. This segment was

retrieved due to its high visual similarity with the query segment. It did not occur among the top retrieval results using text-based similarity only.

Several keyframes of the query segments in which visual similarity degraded the results are shown in Figure 4.9. These keyframes mainly display people and do not contain any specific content or background. Therefore, even an optimum visual similarity calculation can hardly be expected to improve the results in these cases.

Figure 4.9: Examples in which employment of visual similarity decreased the retrieval quality.

Combination with text-based retrieval

In order to combine text-based retrieval with visual similarity, the similarity between pairs of keyframes was used in the presented experiments. Specifically, the maximum similarity was calculated between any keyframe from the query segment and any keyframe from the target segment. However, this strategy is not optimal as single keyframes can be similar to each other accidentally, e.g. both segments can contain black keyframes. Thus, several other strategies were tested in the follow-up experiments. Total visual similarity was also calculated as a sum of visual similarity of the two or three most similar target segment keyframe / query keyframe pairs. Moreover, these two or three keyframes from the most similar pairs can be selected from segments with or without repetition. One keyframe can be thus taken into account either once or multiple times.

Achieved scores are tabulated in Table 4.6. Even though the experiments were performed on the same data as those in the previous section, the baseline system in these experiments used 60-second (instead of 50-second) segments, since it previously performed better in the window-based segmentation, and no metadata and context was utilized. The approach which used similarity between two keyframes (each one was selected from the segment without a repetition)

achieved the most promising results. For the three most similar keyframes, the MAP-bin score was slightly higher but the improvement of the MAP-tol score decreased a bit.

Transcript	Keyframes	MAP-bin	MAP-tol
Subtitles	None	0.146	0.134
Subtitles	Max 1	0.156	0.139
Subtitles	Max 2	**0.162**	**0.145**
Subtitles	Max 3	**0.162**	0.142

Table 4.6: Results for visual similarity calculated using different numbers of keyframes. Best results are in **bold type**.

Additionally, in the case of calculating similarity using two keyframes, only keyframes which were similar enough to their adjacent keyframes were taken into account. This simulated a situation where the selected keyframes are specific representatives of the segments. Similarly, only keyframes which differed enough from adjacent keyframes were taken into account, which simulated a scenario where selected keyframes are distinctive compared to their surroundings. In addition to maximum similarity, average similarity between all pairs of keyframes in query and target segments was also calculated. Each of these approaches increased both MAP-bin and MAP-tol scores, but they did not outperform the maximum similarity between any two selected keyframes referenced in Table 4.6.

Object recognition

Similarity between keyframes was also calculated using the AlexNet CNN visual features (Krizhevsky et al., 2012). These CNN-based similarity scores were provided by the Laboratory of Data Intensive Systems and Applications at Masaryk University. Calculated object similarity was linearly combined with text-based similarity in the same manner as similarity calculated using Feature Signatures (Equation 4.2). Like Feature Signatures, CNN purely improved text-based retrieval, but the improvement was smaller than in the case of Feature Signatures and thus CNN techniques were not applied in the 2015 Video Hyperlinking Task.

Deep neural networks

Deep neural network models recently outperformed “traditional ML models” in numerous applications including image recognition (Ciresan et al., 2012; Szegedy et al., 2015), ASR (Hinton et al., 2012) and machine translation (Wu et al., 2016).

The functional concept of a typical deep convolutional neural network is displayed in Figure 4.10[15].

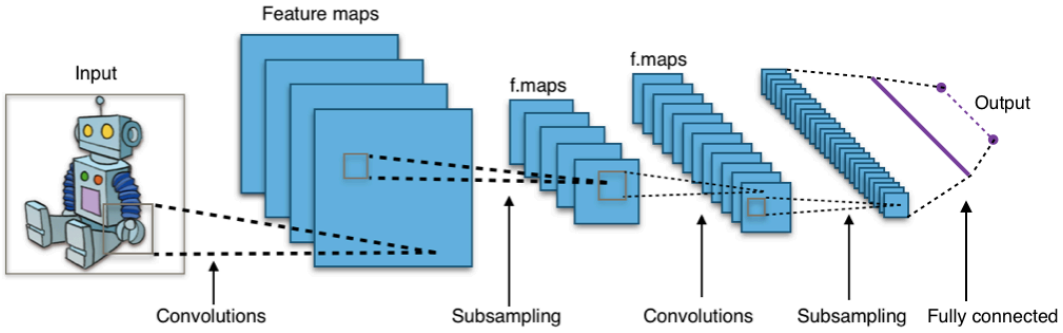


Figure 4.10: Conceptual diagram of a typical convolutional neural network.

Convolutional neural networks consist of sequences of multiple layers. The input image is saved in the *input* layer. A set of filters (masks of pixels with associated values) is then applied to the input image, slid over local regions of pixels and multiplied with image pixel values. Multiplied values are then summed and a *feature map* is created for each filter. The filters eliminate the need to use manually crafted features as in traditional ML. A *rectified linear unit* is then typically applied to the feature maps and an *activation function* (e.g. $\max(0, x)$) is applied elementwise to individual pixels. The *pooling layer* then downsamples the input feature map and reduces its spatial size to eliminate overfitting. Finally, the *fully connected layer* connects the output of the previous layer, consisting of trained features, with the output classes and returns the class scores.

In the object recognition experiments, the final layer was eliminated and the output from the last hidden layer, fc7, of AlexNet was used as the trained features for calculation of visual similarity between query and data keyframes. These 4096-dimensional vectors were organized for efficient similarity searching by a distance-based index (Novák et al., 2015) and this system provided the similarity scores.

Feature Signatures and CNN combination

Both visual similarity calculation models were also combined. First, they were combined linearly with the output of text-based retrieval as in the previous

[15] Image was created by Aphex34 (Own work) [CC BY-SA 4.0 (http://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons

experiments:

$$\begin{aligned} \mathit{FinalScore}(\mathit{segment/query}) &= \mathit{Score}(\mathit{segment/query}) \\ &+ \mathit{Weight1} * \mathit{FeaturesSignaturesSimilarity}(\mathit{segment/query}) \\ &+ \mathit{Weight2} * \mathit{CNNSimilarity}(\mathit{segment/query}). \end{aligned} \quad (4.3)$$

In order to acquire comparable scores, maxmin and mean normalizations were applied to both visual similarity scores and both weights were jointly tuned. However, this approach did not outperform the combination of text-based retrieval and Feature Signatures.

Scores were also combined using re-ranking. CNN techniques were first applied and combined with text-based search in order to retrieve similar objects. The top retrieved results (10, 20, 50 and 100 items) were then re-ranked using Feature Signatures in order to sort target keyframes according to their "overall" similarity with the query keyframe. Although this approach gave a small improvement, it again did not outperform results achieved by basic linear combinations of text and Feature Signatures.

Concept detection

Convolutional networks were also used to generate textual descriptions of the concepts depicted in the keyframes. Concepts were created by the image annotation system used in the ImageCLEF 2014 Task (Budíková et al., 2014) and provided by the Laboratory of Data Intensive Systems and Applications at Masaryk University. This system first retrieves images similar to a given query image, or keyframe in our case, from annotated Profiset dataset (Budíkova et al., 2011) using CNN. The word descriptions of the most similar images are then used for text analysis where the word semantic relationships are exploited to create a description of the query image. A description is created for each keyframe as a list of concepts (e.g. *people*, *indoors*, *young*, *two*, *canadian*, *plant*, *macro*) with a confidence score assigned to each concept.

Detected concept descriptions were appended to the created text-based queries and each word occurrence from the content description (from all keyframes inside the query segment) and each word occurrence from the query were used to create new queries. When all detected concepts of all query keyframes were used, both scores decreased since the concepts then contained excess noise. Using only concepts with higher confidence scores and concepts with a restricted number of occurrences (occurring up to two or three times in the query segment) improved the results but the scores were still lower than the baseline. However, assigning concept confidence scores as query term weights increased both the scores. The

highest MAP-tol result was achieved using a combination of concept term weighting and a restriction of the number of occurrences of each concept in the query (up to six occurrences). The results achieved for query expansion using concepts are shown in Table 4.7. Due to the nature of concepts, it can be expected that they can be even more useful for expansion of data segments since concept terms can relatively often be expected to occur in queries but not in speech. However, this approach was not pursued since it is somewhat computationally demanding.

Transcript	Concepts	Max Concepts	MAP-bin	MAP-tol
Subtitles	None	–	0.233	0.138
Subtitles	Yes	–	**0.241**	0.160
Subtitles	Yes	5	0.237	**0.164**

Table 4.7: Query Expansion by visual concepts with and without restriction of the maximum concept occurrences. Concepts are weighted by their confidence scores in both cases. Best results are in **bold type**.

Lists of concepts detected for each keyframe were provided by the task organizers and were created by Leuven (Tommasi et al., 2014) and Oxford (Chatfield et al., 2015) Universities. Both teams provided data with 1537 image net (Deng et al., 2009) concepts which correspond to WordNet synsets. However, we only used concepts provided by the Masaryk University, since they performed well in the ImageCLEF 2014 Scalable Concept Image Annotation Task (Budíková et al., 2014).

Face recognition

In addition to the recognition of settings and particular objects, experiments with face recognition were also performed. Face descriptors were provided by the Center for Machine Perception at Czech Technical University in Prague. Faces were detected by a commercial multi-view face detector[16]. Each face was geometrically aligned with a canonical pose by automatically extracted facial features. Then a compact vectorial SIMILE descriptor (Kumar et al., 2009) was calculated. A set of face descriptors representing persons' identities were thus available for each keyframe in which faces were detected. Faces were then compared by L2 distance on calculated descriptors.

In this way, it was possible to discern the presence of specific persons in the dataset based on the query keyframes. Detected similarity was first linearly combined with text-based similarity in the same way with visual similarity. This

[16] Eyedea Recognition Ltd. http://www.eyedea.cz

combination very slightly increased the MAP-bin score from 0.146 to 0.147 but the MAP-tol score dropped from 0.134 to 0.132. Additionally, face similarity was used for re-ranking the top 1000 text-based retrieval results. In this case, both scores were improved. MAP-bin increased from 0.205 to 0.209 in the case when only faces with similarity scores higher than a given threshold were used. The MAP-tol score increased from 0.116 to 0.128 when all available faces were used. Due to unstable improvements achieved with the training data, this approach was not applied to the test data. Employing recent deep neural network descriptors that are more powerful, e.g. (Parkhi et al., 2015), can be expected to further improve the performance. Face recognition can be also expected to be helpful in very specific user scenarios.

4.3.2 2015 Video Hyperlinking methods comparison

The 2014 MediaEval SH Task and the 2015 TRECVID Video Hyperlinking Task shared the same dataset but differed by the set of queries used for evaluation. The test set used for evaluation of the SH Task was available for training purposes in the Video Hyperlinking Task. Therefore, approaches which achieved the most promising results in the 2014 MediaEval SH Task experiments and in described follow-up experiments were submitted to the 2015 TRECVID Video Hyperlinking Task. The submitted approaches are presented in this section.

Baseline run

The core of the system was similar to the set-up used in the 2014 MediaEval SH Task experiments. All recordings were segmented into 60-second long passages with new passages being created every 10 seconds. Transcripts of the passages created from available subtitles were concatenated with corresponding metadata of the video file. A title, a description and information about the broadcast channel, which was mined from the filename, were used. A context of 20 seconds was also utilized.

The query segments were expanded by audio information contained in each segment as described in Section 4.2.6. Sub-segments were submitted to the Doreso service to retrieve music composition titles and artist names which were then concatenated with the query segment. Music was only detected in 7 out of 135 test queries (e.g. query 96 contained *Cassava* by *Triclops!* and *Safari* by *John Barry*, and query 69 contained *Something To Talk About* by *Badly Drawn Boy*).

Each query was further expanded by visual concepts contained in the source video segment as described in Section 4.3.1. Concepts were again provided by the Laboratory of Data Intensive Systems and Applications at Masaryk Univer-

sity. Each concept had an associated confidence score that was used to weight individual terms in the concatenated query. Each concept was only used if it occurred in less than 7 segment keyframes.

Tuning the baseline

The highest improvement on the training data was achieved by combining the baseline data with information on whether the data and query videos came from the same TV series. This run is further referred to as the **Series** run. The information about the TV serie available in the metadata was used to precalculate a "series weight" for each video segment. The series weight was set to 0.13 if the query video and data video were from the same TV series; otherwise the weight was set to -0.15. These weights were calculated using the training data by employing precision of the results retrieved from the same and different TV series. The average precision of the results retrieved from the same TV series is 0.59, and the average precision being in a the different TV series is 0.44. Weights were then calculated for each query on the test data and they were linearly combined with the top 1000 retrieved results.

Similar to series weights, experiments were also performed with "time differential weights" determined by time differences between the data and query video dates. On the training data, we confirm that the precision of videos broadcast up to one day from the query video (0.59) is higher than the average precision values (0.34) and also higher than the Precision of the videos broadcast up to one week from the query video (0.35). Also, videos broadcast before the query segment have slightly higher average precision values (0.36) than videos broadcast after the query segment (0.32). Time differential weights were combined with the baseline system in the same way as TV series weights were used. However, the approach which favored videos from the same TV series achieved a greater improvement with the training data and were thus not submitted to the official evaluation.

Another run submitted to the task combined the baseline and visual similarity and this run is further referred to as the **FS** run. In this run, the baseline was expanded by the visual similarity between the query segment and data segments which was calculated using Feature Signatures as described in Section 4.3.1.

Finally, the last submitted run is a combination of the FS and Series runs. This run is further referred to as the **FSRerank** run. The top 1000 results returned by the FS system were linearly combined with the same weights as those used in the Series run and re-ranked accordingly. In the FS and FSRerank runs no segments were retrieved for the query 118. For this query, the answers

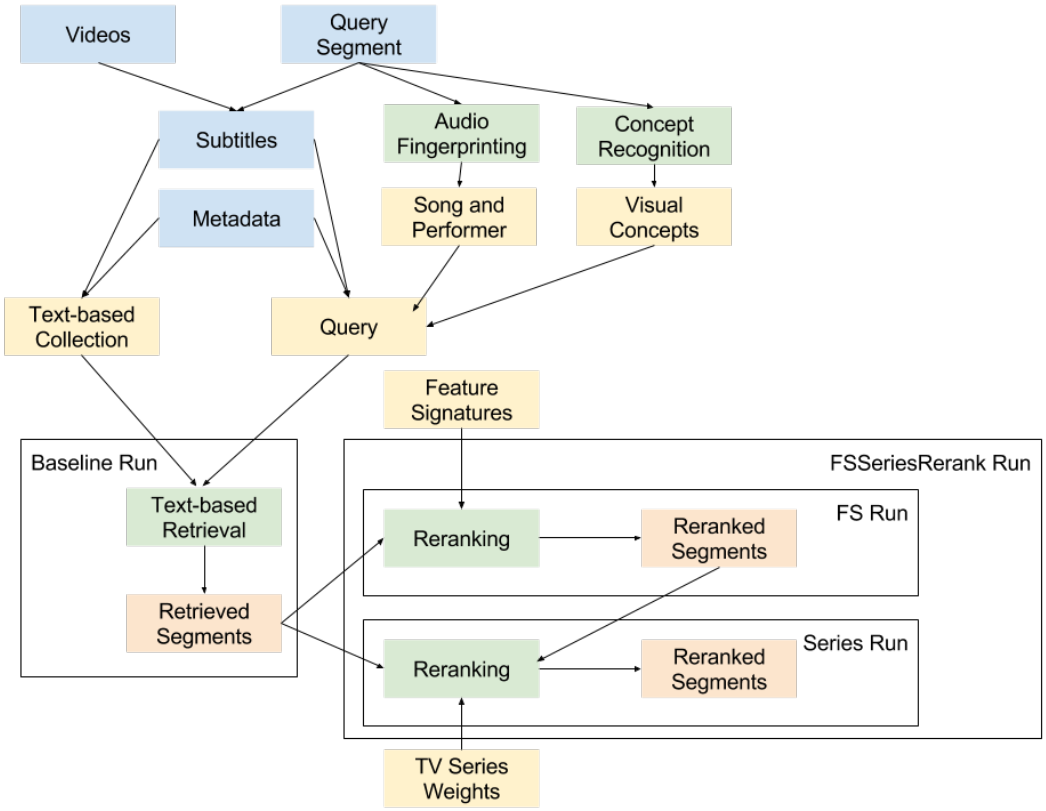


Figure 4.11: Retrieval processing system - strategy diagram.

from the baseline run were used. All experiments were tuned for the highest MAP-tol measure.

TRECVID 2015 Video Hyperlinking Task results

The strategy diagram of the retrieval approaches is displayed in Figure 4.11. Performance comparisons are reported in Table 4.8.

Run Num.	Run Name	MAP-bin	MAP-tol
1	Series	0.144	0.113
2	FSSeriesRerank	0.109	0.084
3	FS	**0.156**	**0.123**
4	Baseline	0.154	0.121

Table 4.8: Performance comparison of results submitted to the 2015 Video Hyperlinking Task. Best results for each performance measure are in **bold type**.

The Series run significantly outperforms the FSRerank run. Both these runs are significantly inferior to the baseline run in terms of both reported performance

measures. The FS run outperforms the baseline run. Although it is not significantly better in terms of the reported scores, testing confirms that employment of visual similarity consistently improves video retrieval based on information contained in subtitles. Weighting using information about the same TV series did not cause the associated system to outperform the baseline run in any case; the baseline run is significantly better in terms of both scores despite its high improvement on the training data.

4.3.3 Conclusion

In this section, we describe our experiments with employing visual information mined from videos in hyperlinking. We confirm that visual information is helpful for multimedia retrieval, but the process of mining visual information needs to be adapted to the type of the archive and to what kind of information is important for the archive users. We tested usage of Feature Signatures convenient for background and setting recognition, CNN for more detailed recognition of objects in the keyframes and text descriptions of the keyframe concepts and face recognition for acquiring information about presence of specific persons in the video. We also tested different combinations of these visual descriptors. In hyperlinking of TV programmes, Feature Signatures performed especially well in our experiments and constantly improved results. We also provide analysis of the cases in which Feature Signatures are helpful and the cases in which they cannot be expected to improve the results. Even though that some of the tested methods did not perform well in our setting, they are still expected to be helpful for specific archives and user needs.

5

Anchor selection

The previous chapter dealt with automatic methods for retrieving segments similar to the predefined source segment. It was assumed that the source segment is either defined by the owner, producer or user of the archive. However, this approach requires additional physical work. Also, even though user-defined anchors provide greater flexibility, since a user can choose any segment of interest, all calculations required for the retrieval need to be executed quickly. Because the high processing time requirements associated with processing large multimedia archives, some retrieval methods may take excessive time and others may even be impossible to run online. This chapter addresses the excessive processing time problem by covering automatic anchor selection which incorporates predefinition of anchoring segments so that the time demanding processing can be run in advance.

Automatic definition of anchors and targets has been investigated, for example, in the Wikification project (Mihalcea and Csomai, 2007). Here the important keyphrases of texts were automatically detected and linked with appropriate Wikipedia articles. Problems associated with automatic anchor selection were more precisely defined and explored in the Anchoring sub-task of the SAVA Task organized at the 2015 MediaEval Benchmark. The anchoring segments should be in some way noteworthy for the archive users to stimulate additional interest in finding more information about them. In a wider sense, anchor selection is related to issues associated with automatic selection of the most representative or interesting video segment and video summarization. Anchor selection described in this chapter assumes that anchors are segments most appropriate for further application of hyperlinking. Thus, the user can easily browse the archive using the links between anchoring segments and related data segments to find information about related topics (see Figure 5.1).

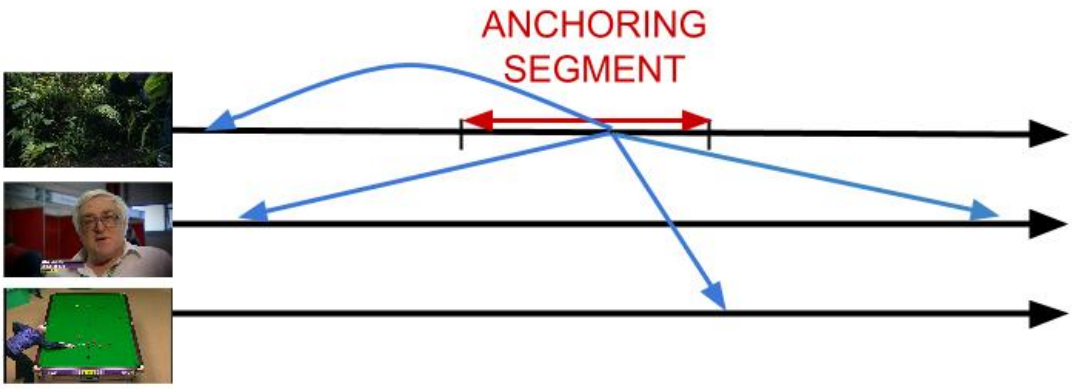


Figure 5.1: Hyperlinking of anchoring and data segments.

Chapter overview

This chapter describes various approaches to automatic detection of anchoring segments on which hyperlinking approaches described in the previous chapter can be further applied. Experiments described in this chapter were performed in the 2015 MediaEval SAVA Task and have been previously published in the benchmark report (Galuščáková and Pecina, 2015).

Data and Evaluation

The Anchoring sub-task uses BBC 2015 SAVA set described in Section 2.1. This dataset is a subset of the BBC 2014 SH dataset: 72 videos from the BBC 2014 SH were selected for the Anchoring sub-task, 38 programmes were used for training and 34 programmes for testing. 90 anchoring segments manually marked in the training data were available for training purposes, but this list was not exhaustive.

The Anchoring sub-task was evaluated using the crowdsourcing campaign. The top 25 results returned by each task participant were manually marked as correct or incorrect. Partially overlapping segments returned by the participants were joined before the annotation. Results were then evaluated using Precision at 10, Recall and MRR scores. Retrieved segments were judged as being relevant if they overlapped any relevant segment. Details of the task, data, and evaluation process are given in the task description (Eskevich et al., 2015).

5.1 Systems description

Approaches used for automatic selection of anchoring segments are based on different assumptions. An approach by Vliegendhart et al. (2015) is based on

social media. Since anchoring segments are intended to be segments for which users may require additional information, Vliegendhart et al. (2015) use Twitter to identify topics on which people have questions. For each shot they first define a set of keyphrases which occur in it based on its associated subtitles. They then find the number of people who asked a question regarding this keyphrase using Twitter. Keyphrases can be further weighted by their popularity. Simon and Gravier (2015) use a hierarchical structure of topically focused fragments of video based on word burstiness. Anchoring segments are then selected from these predefined segments using lexical cohesion.

Two other approaches are presented in this chapter, one based on metadata, the other based on frequencies of occurrence of proper names and numbers contained in the segments (Galuščáková and Pecina, 2015). The basic assumption in these approaches is that since the anchoring segments are intended to be further used as a basis for subsequent hyperlinking to other related video segments, anchoring segments are mainly intended to be useful and informative for the users of the archive. Both approaches proved to be helpful for different aspects of the anchoring problem: the segments which contain a large number of proper names and numbers are interesting to the users, while the segments most similar to the video description provide high information content.

Selection using Passage Retrieval

Both approaches make use of subtitles available for each program. All videos were first segmented into shorter passages, which become candidates for anchoring segments. The shorter passages were either created regularly every 10 seconds (regular window-based segmentation) with a segment length of 60 seconds or by using machine learning feature-based methods which automatically detect probable segment ends (ML segmentation). Both methods are described in Section 3.3.2. Afterwards, created segments were sorted according their probability of being the segment of interest (anchoring segment) for the archive users.

In the **first** method, manually-created metadata available for each program were converted into queries. The metadata consist of the program name and a short program description. An IR system was then used to sort the created segments according to their similarity with metadata describing the program. Since the metadata provide a good description of the content of the recording, the most similar segments are expected to also contain relevant content and adequately describe the video recording.

In the **second** method, the segments were sorted according to the frequency of occurrence of numbers (words containing at least one digit) and proper names (words containing an upper case character while the previous word does not end

with a dot) contained in each segment. Segments containing more proper names and numbers are expected to contain some kind of specific information.

The Terrier IR Framework, its implementation of the Hiemstra LM with its parameter set to 0.35, the Porter stemmer and Terrier's stopwords list were used in all experiments. IR was also used for sorting the segments based on the proper name and number occurrence frequencies. Since the ongoing plan is to combine the score from the retrieval based on metadata with the segment preference given by proper name and number occurrence frequencies, the retrieval was first run in the same way as in the case when only metadata were used. The weights corresponding to the frequency of number and name occurrences were precalculated. Finally, the precalculated weights were linearly combined with the weight acquired as the output of the IR. The combined weight was then set to greatly favor the ordering given by proper name and number occurrence frequencies.

5.2 Results

Officially evaluated experimental results are displayed in Table 5.1. Machine learning-based segmentation proved to increase the precision compared to regular segmentation. Due to a large number of overlapping retrieved segments, which is even larger for machine learning-based segmentation, this level of Precision measure confirms the quality of the selection of the top-retrieved results. The regular segmentation without any proper name and number preferences achieved the overall highest MRR score. This confirms the assumption that highly ranked segments most similar to the metadata are also the most informative ones.

Segmentation	Preference	P10	Recall	MRR
Reg	—	0.279	0.251	**0.926**
ML	—	0.303	0.250	0.903
Reg	Numbers	0.279	0.243	0.891
Reg	Names	**0.312**	0.270	0.861
Reg	Numbers and Names	**0.312**	**0.271**	0.834

Table 5.1: Results officially achieved in the SAVA Task: a comparison of segmentation strategies and a preference based on frequencies of occurrence of proper names and numbers contained in the segments. Best results for each measure are in **bold type**.

The overall highest Precision and Recall scores are achieved when segments contain the largest number of proper names and numbers. This confirms the assumption that when a segment contains a large number of proper names and

numbers, it is likely to be interesting to the archive user. The same Precision was also achieved when only proper names were used. Difference in Recall is minor and MRR is even higher in such cases.

Future plans

Overall, high MRR scores achieved by conversion of metadata into queries and their utilization for retrieving the most informative parts of the document proved this approach to be convenient for anchor selection. Achieved Precision and Recall numbers show that the combination of this approach with the preference of the segment based on the occurrence frequencies of proper names and numbers may further improve the results. However, this combination would need some detailed tuning. The detection of proper names should be solved properly using the named entity recognition. Utilization of other information such as occurrence frequency of the content words and mentions of places and persons would also need further exploration.

5.2.1 Post-processing

Since described achieved scores were lower than those achieved by two other teams participating in the shared task, results were re-examined. Because the retrieved segments were not filtered out, the top results contained many partially overlapping segments. In the evaluation, each relevant segment was only used once and many of these overlapping segments were thus considered to be irrelevant. Therefore, results were post-processed after the official evaluation, partially overlapping segments were filtered out and segments were evaluated with respect to all top submitted results. Some of the top results from the post-processed runs were unjudged as they did not previously occur among the top results which were judged. On average, one to two segments from the top 10 retrieved segments were not judged and calculated scores may thus be slightly smaller than real scores. Achieved scores are displayed in Table 5.2.

All achieved scores are higher than the officially achieved scores. Six segments out of the top 10 results are marked as possible anchoring segments by human evaluators on average and around half of all anchoring segments marked by human annotators were found. Moreover, Recall and MRR outperform the overall best results achieved in the task. In general, differences between scores achieved by examined methods are small, especially in the case of Precision. The biggest difference is in Recall which is slightly lower for ML-based segmentation. Highest MRR scores are achieved when no preference or preference based only on proper names is used.

Segmentation	Preference	P10	Recall	MRR
Reg	—	**0.603**	0.546	**0.949**
ML	—	0.600	0.492	0.934
Reg	Numbers	0.591	**0.549**	0.929
Reg	Names	**0.603**	0.546	**0.949**
Reg	Numbers and Names	0.591	**0.549**	0.929

Table 5.2: Scores of filtered results achieved in the SAVA Task: a comparison of segmentation strategies and a preference based on frequencies of occurrence of proper names and numbers contained in the segments. Best results for each measure are in **bold type**.

5.3 Conclusion

This section describes a problem of an automatic selection of anchoring segments which are segments convenient to be origin segments for hyperlinking. Automatic selection of the anchoring segment in multimedia is a relatively new problem. However, solving this task is important for examining alternative approaches to navigation in multimedia archives. Anchor selection is important for further application of hyperlinking and it can also bring additional information to users what can make navigation in archives even easier. We provide our experiments with automatic selection of anchoring segments based on metadata and on the number of occurrences of proper names and numbers. Our experiments did not perform well in the official evaluation of the 2015 SAVA Task, but we achieved competitive results after the overlapping segments were filtered out.

6

User interfaces for video search

This chapter describes SHAMUS (UFAL Search and Hyperlinking Multimedia System), a complex open source system which provides easy search and navigation capabilities for multimedia archives. The system provides a graphical user interface for three components described in the previous chapters. The **search** component provides text-based search in multimedia archives, the **anchoring** component determines the most important segments of videos, and segments topically related to the anchoring segments are retrieved by the **hyperlinking** component. A demo of the system working with a TED Talks dataset is available online[1]. The system is primarily intended for multimedia professionals and those doing exploratory searches for which content-based search methods are more appropriate. However, the system can also be used for entertainment purposes.

First, a general overview of multimedia search and hyperlinking user interfaces is discussed. The SHAMUS user interface, all components of the SHAMUS system, and the TED Talks dataset used in demo are then described. SHAMUS has been previously described in the demo paper (Galuščáková et al., 2016) published at the European Conference on Information Retrieval.

6.1 Overview of video browsing and retrieval user interfaces

According to Schoeffmann et al. (2010a), graphical user interfaces of video browsing and retrieval systems serve as a mediator between the available dataset and the user. These interfaces thus enable users to approach the archive, browse it and search for documents relevant to their needs. Schoeffmann et al. (2010a) also

[1] `http://ufal.mff.cuni.cz/shamus`

identify the main problems with which search and browsing user interfaces need to deal with. These are "how to query video data, how to effectively retrieve results and how to effectively present the content". Though these problems also arise in text retrieval user interfaces, they are much more pronounced in video browsing and retrieval systems.

Queries are traditionally textual keywords, visual examples and concepts (Snoek et al., 2007). Textual keywords is the most frequently used method, well known to most users. However, it can be insufficient for image and video browsing and retrieval interfaces. In a query-by-example paradigm, the user typically provides an image as the input query. Some interfaces allow users to sketch the query (James and Collomosse, 2014) or choose requested colors (Kruliš et al., 2013). Query-by-concept (Snoek and Worring, 2009) allows user to describe the required output directly by semantic concept which is then mapped to low-level features (Schoeffmann et al., 2010a). The type of query is also closely connected to the retrieval method. Based on the retrieval method, systems need to deal with the problems of how to represent the query and documents, how to calculate a distance between these representations and which documents to retrieve based on the calculated distance. Some of these retrieval methods were described in Chapter 2 and Chapter 4.

Visualization of the results should enable users to understand retrieved content (Schoeffmann et al., 2010a) and ideally enable them to quickly decide if the results are relevant to their information needs. Various methods are used to describe the video search results. Results are reported using a textual snippet or video description, a single image (Li et al., 2013; Moumtzidou et al., 2015), a series of frames (e.g. most representative frames, clusters of frames, uniformly sampled frames) (Heesch et al., 2004; Lokoč et al., 2015) and by using various video surrogates (Campanella et al., 2005; Goëau et al., 2007; Schoeffmann et al., 2010b).

Browsers also differ with respect to potential users and tasks. Efficient navigation methods are especially needed by professional archive users (e.g. historians, librarians, content producers), who are searching for particular information and need to explore topics in detail. These users are also likely to use more complex search methods. Quality and effectiveness of video browsing interfaces, in terms of how quickly the user is able to locate a required document was explored in VideOlympics (Snoek et al., 2008) which is now known as Video Browser Showdown[2] (Schoeffmann, 2014). Participating teams test their systems on a moderately large video archive and try to interactively find a short video clip as fast as possible in front of an audience.

[2] hhttp://www.videobrowsershowdown.org

Interfaces for hyperlinking

Video search interfaces are most typically based on a *Query and Search* paradigm. This method requires a user to name the main concept of the query, which does not have to be obvious or may not even exist in some cases. Therefore, some systems (Gordon and Domeshek, 2001) use a *Zoom and Browse* method instead. These systems do not require the user to formulate a query but they allow filtering of the content and browsing the archive. The Zoom and Browse methods thus include hyperlinking.

The Hyper Video Browser (Eskevich et al., 2015) provides an interface for text-based search and hyperlinking of video segments. Similar to the presented system, the search is based on metadata and subtitles. In addition, the system also utilizes detected visual concepts and thus also enables searching using visual information. Distance between query and detected visual concepts is calculated using WordNet distance. Detected concepts are also used to create hyperlinking queries. An archive can also be browsed using a tag cloud which consists of metadata terms and visual concepts. Visual information (detection of shots, visual concept detection and optical character recognition) was also used in the segmentation process. In contrast to the presented solution, the system does not automatically mark interesting anchoring segments. Hyperlinks within the currently played video are displayed on the playback timeline instead.

6.2 System components

In the SHAMUS system, search, described in Chapter 2, is applied to available subtitles. In addition to this text-based search, SHAMUS also provides hyperlinking, described in Chapter 4, which recommends links to other related segments on the fly. In order to make the system more adaptable to different datasets, both components use only textual information. Since the size of the TED dataset is substantially smaller than the size of the BBC 2014 SH dataset which was used for system tuning, using full transcription of the query segment was ineffective in the demo. Therefore, only the 20 most frequent words lying inside the query segment are used to formulate the hyperlinking query. Moreover, the most common words are filtered out using the stopwords list. A number of used words was tuned using the BBC 2014 SH dataset. Approaches which only use the first words of a segment and words with the highest TF IDF were also tried, but they were outperformed by the most common words.

SHAMUS also automatically suggests the most informative segments using the anchoring component, described in Chapter 5. Segments which are most similar to the document metadata description are used as the most informative

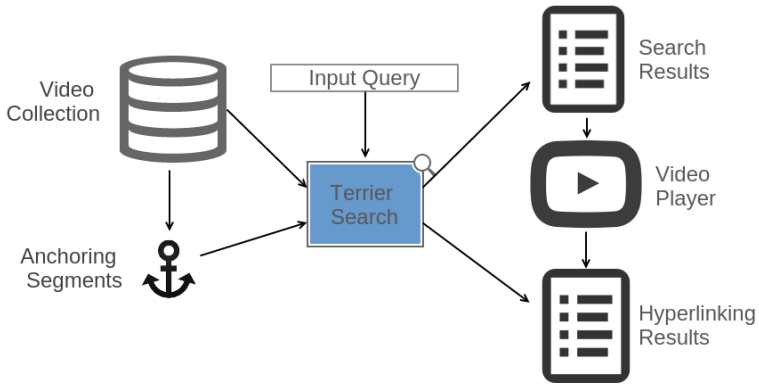


Figure 6.1: SHAMUS - strategy diagram of the system.

ones. The list of anchoring segments is pre-generated in advance for each video and does not vary depending on the query. Hyperlinking is then run on detected anchoring segments. The list of related segments is automatically regenerated on the fly when a new anchoring segment begins to play. All components work with 60-second long segments created every 10 seconds and "overlap removal" filtering is applied to the retrieved segments. IR framework Terrier is used in all system components. The strategy diagram of the SHAMUS system is displayed in Figure 6.1.

6.2.1 TED Talks dataset

The SHAMUS demo interface works with a dataset of 1219 TED Talks. A list of talks available in the TED dataset (Pappas and Popescu-Belis, 2013) was used for this purpose. Subtitles and videos of each talk from this list were downloaded directly from the TED website. The TED dataset was also used for mining metadata such as title and description. The dataset consists of talks presented and recorded at TED conferences and topics include technology, entertainment, design, business, science and global issues. Most of the talks are up to 20 minutes long. All subtitles were created manually by volunteer transcribers. Talks are available under a Creative Commons non-commercial license. Subtitles and videos are not published with the system but the archive contains filenames, scripts for downloading, pre-processing and segmentation of all data needed for running the system. The published archive contains metadata which are stored in a database and a list of anchoring segments from the TED dataset with transcribed segment beginnings. SHAMUS can also be easily adaptable to work with any dataset of videos for which subtitles or automatic transcripts and metadata descriptions are available.

6.3 SHAMUS user interface

The SHAMUS user interface consists of three main website sections. The first one services textual query inputs (see Figure 6.2). The second section displays the results of the search, including the metadata, transcript of the beginning of the retrieved segment and playback time of this segment (see Figure 6.3). All segments from a single video file are grouped together and displayed in a list. The video with its title, description, source, marked anchoring segments, and list of related segments are displayed at the third section (see Figure 6.4).

Figure 6.2: Entry website of the SHAMUS system with text-based search.

The JWPlayer[3] is used for the video playback. The anchoring segments are marked as individual chapters of the video. The transcript of the beginning of each segment is retrieved and used instead of the chapter name. Users can thus view the most important segments of the video without the need to navigate through the video. A list of the three most related segments is displayed on the right side of the video player. This list is re-generated each time any new anchoring segment begins. The title, file description and playback time of the beginning of each segment are displayed in the list. When a user clicks on any of these segments, video playback is reloaded and the linked video starts to play from the playback time displayed in the list. The links can also exist within a

[3] `http://www.jwplayer.com`

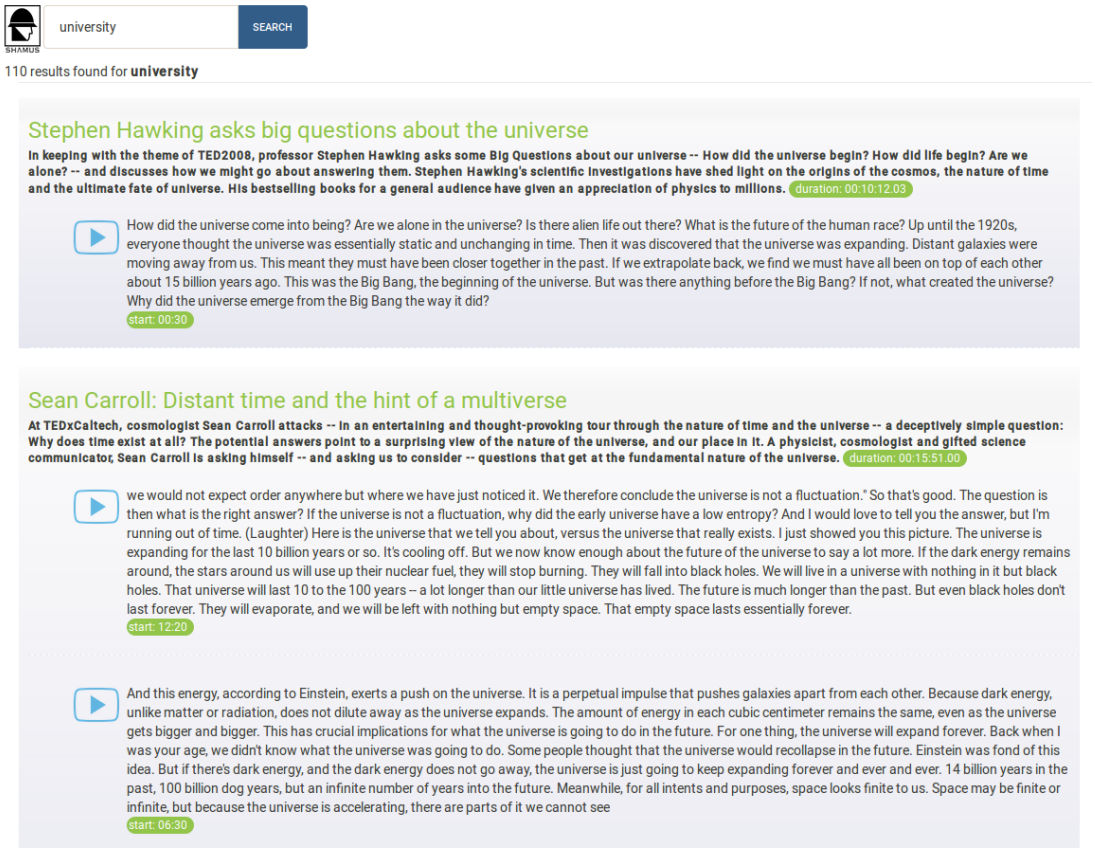


Figure 6.3: List of segments retrieved from the text-based search.

Figure 6.4: Video player with the detected most informative anchoring segments and recommended links to related video segments listed at right.

single video. The displayed links can thus be used for navigation within the video and also throughout the entire archive.

Summary

This thesis overviews and compares various approaches for retrieval from audio and audio-visual documents. It focuses on two types of retrieval – *search* which processes textual queries (Chapter 2) and *hyperlinking* which identifies segments semantically similar to a query segment (Chapter 4). The goal for both tasks is to retrieve precise relevant segments of audio and video documents and thus make navigation in audio-visual archives faster, more precise and more convenient for users. The described approaches only use content of the recordings. Textual information mined from subtitles and ASR transcripts is combined with audio (Section 4.2) and visual (Section 4.3) information and available metadata. The described approaches are tuned and tested on several benchmark tasks organized at MediaEval (Eskevich et al., 2012b, 2013b, 2014b, 2015) and TRECVID (Over et al., 2015).

Segmentation used in applied Passage Retrieval (Chapter 3) methods substantially influences search quality levels of text-based retrieval and hyperlinking processes. Window-based segmentation into passages of equal length performs well and is robust. However, segment length and segment shift are crucial for this type of segmentation and overlapping segments occurring in the list of retrieved segments need to be properly processed. Despite the good results achieved using window-based segmentation, the system can be further improved using segmentation based on ML-based methods which detect probable segment boundaries. Moreover, Passage Retrieval methods have been confirmed to outperform retrieval of full documents.

The presented retrieval methods rely heavily on textual information from subtitles and transcripts, and both search and hyperlinking tasks are first transformed into text retrieval sub-tasks. However, automatic transcripts are influenced by problems inherent in ASR systems, predominantly a restricted vocabulary and reliability of the output. Restricted vocabulary can be adequately addressed either by expanding data and queries by metadata information or by

combining different transcripts. Transcript reliability can be improved by using only words having the highest confidence levels as opposed to all available word variants returned by the ASR system. Audio information may also be employed as calculated acoustic similarity and combined with scores from text retrieval. It is also possible to use detected music from the query segment to expand the textual query. Even though these two methods need further study they can be especially helpful for specific retrieval needs.

Improvements in text-based retrieval may also be achieved through the use of visual information. Scores from textual retrieval are combined with visual similarity between segment scores calculated as the distance between representative keyframes represented by Feature Signatures. This descriptor is especially helpful in retrieving segments with similar settings and backgrounds. Other visual processing methods (similarity calculated using deep neural networks, concept detection and face recognition) improve results achieved on training data and need additional comprehensive testing.

The last section presents our video search and navigation, open source interface, SHAMUS (Chapter 6). This hyperlink navigated interface supports text-based video searches and automatic detection of probable anchoring segments (Chapter 5) . These segments are selected as being most similar to the video metadata description and are thus intended to be highly informative to the user. A demo interface which works with a dataset of TED Talks (Galuščáková et al., 2016) is available online.

List of Figures

List of Tables

Abbreviations

AP Average Precision. 14

ASR Automatic Speech Recognition. 3, 8, 24, 45, 49–52, 58–60, 73–75, 78, 88, 111, 112

CG Cumulative Gain. 15

CNN Convolutional Neural Networks. 81, 82, 88–90, 95

DCG Discounted Cumulative Gain. 15

GAP Generalized Average Precision. 18, 19

IR Information Retrieval. vii, 2, 3, 7, 8, 10, 14, 18, 24, 25, 27, 30, 34, 39, 41, 42, 44, 45, 49–52, 71, 73, 99, 100, 106

LDA Latent Dirichlet Allocation. 47, 72

LM Language Model. 26–30, 32, 33, 35, 36, 39, 50, 54, 73, 100, 113, 114, 116, 117

MAiSP Mean Average interpolated Segment Precision. 18, 20, 71

MAP Mean Average Precision. 14, 19, 76

MAP-bin Binned Relevance. 18–20, 43, 62, 63, 68, 71, 77, 78, 80, 85, 88, 91, 92, 94, 113

MAP-tol Tolerance to Irrelevance. 18–20, 43, 62, 63, 68, 71, 77, 78, 80, 85, 88, 91, 92, 94, 113

ABBREVIATIONS

MASP Mean Average Segment Precision. 18, 19, 28, 30, 32–35, 37, 54, 55, 113

mGAP Mean Generalized Average Precision. 18, 19, 28, 30–35, 37, 52–54, 58–61, 113, 114

ML Machine Learning. 53, 55, 56, 59–63, 88, 89, 99–102, 111

MRR Mean Reciprocal Rank. 14, 18, 20, 37, 52, 53, 59–61, 98, 100–102

MRRw Mean Reciprocal Rank Window. 18, 28, 30–35, 37, 38, 52–54, 58–61, 114, 117

NDCG Normalized Discounted Cumulative Gain. 14, 15

P10 Precision at 10. 14, 100, 102

ROC Receiver Operation Characteristics. 14, 15

ROO Overlap Removal. 31–36, 50, 52, 54, 62, 85, 113, 114, 117

RR Reciprocal Rank. 14

SAVA Search and Anchoring in Video Archives. 5, 10–12, 16, 17, 97, 98, 100, 102, 118

SH Search and Hyperlinking. vii, viii, 5, 10–13, 16, 17, 20–23, 25, 36, 41–43, 50, 52–54, 58–63, 66–70, 73–76, 80, 82, 85, 86, 92, 98, 105, 113, 114, 116, 117

SSSS Similar Segments in Social Speech. 5, 10, 13, 16, 20, 42, 50–52, 54–56, 58–62, 67, 70, 72, 114, 116, 117

WER Word Error Rate. 75–77, 117

Bibliography

Ah-Pine, J., Csurka, G., and Clinchant, S. (2015). Unsupervised Visual and Textual Information Fusion in CBMIR Using Graph-Based Methods. *ACM Transactions on Information Systems*, 33(2):9:1–9:31. ACM.

Allan, J. (1995). Relevance Feedback with Too Much Data. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '95*, pages 337–343, Seattle, WA, USA.

Aly, R., Eskevich, M., Ordelman, R., and Jones, G. J. F. (2013). Adapting Binary Information Retrieval Evaluation Metrics for Segment-based Retrieval Tasks. *CoRR*, abs/1312.1913.

Andriole, S. (2015). Unstructured Data: The Other Side of Analytics. http://www.forbes.com/sites/steveandriole/2015/03/05/the-other-side-of-analytics.

Ballantine, J. (2004). Topic Segmentation in Spoken Dialogue. Master's thesis, Macquarie University.

Baluja, S., Seth, R., Sivakumar, D., Jing, Y., Yagnik, J., Kumar, S., Ravichandran, D., and Aly, M. (2008). Video Suggestion and Discovery for Youtube: Taking Random Walks Through the View Graph. In *Proceedings of the 17th International Conference on World Wide Web, WWW '08*, pages 895–904, Beijing, China.

Barnett, J., Anderson, S., Broglio, J., Singh, M., Hudson, R., and Kuo, S. W. (1997). Experiments in Spoken Queries for Document Retrieval. In *Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997*, volume 3, pages 1323–1326, Rhodes, Greece.

Beecks, C., Lokoč, J., Seidl, T., and Skopal, T. (2011). Indexing the Signature Quadratic Form Distance for Efficient Content-based Multimedia Retrieval. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval, ICMR '11*, pages 24:1–24:8, Trento, Italy.

Beeferman, D., Berger, A., and Lafferty, J. (1999). Statistical Models for Text Segmentation. *Machine Learning*, 34(1-3):177–210. Kluwer Academic Publishers.

Bendersky, M., Garcia-Pueyo, L., Harmsen, J., Josifovski, V., and Lepikhin, D. (2014). Up Next: Retrieval Methods for Large Scale Related Video Suggestion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1769–1778, New York, NY, USA.

Berners-Lee, T. and Cailliau, R. (1990). WorldWideWeb: Proposal for a HyperText Project. Proposal, CERN.

Bhatt, C. A., Pappas, N., Habibi, M., and Popescu-Belis, A. (2014). Multimodal Reranking of Content-based Recommendations for Hyperlinking Video Snippets. In *Proceedings of International Conference on Multimedia Retrieval*, pages 225–232, Glasgow, UK.

Blažek, A., Lokoč, J., and Skopal, T. (2014). Video Retrieval with Feature Signature Sketches. In *Similarity Search and Applications - 7th International Conference, SISAP 2014*, pages 25–36, Los Cabos, Mexico.

Blažek, A., Lokoč, J., Matzner, F., and Skopal, T. (2015). Enhanced Signature-Based Video Browser. In *MultiMedia Modeling: 21st International Conference, MMM 2015, Part II*, pages 243–248, Sydney, Australia.

Blei, D. M. and Moreno, P. J. (2001). Topic Segmentation with an Aspect Hidden Markov Model. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 343–348, New Orleans, LA, USA.

Boreczky, J. S. and Rowe, L. A. (1996). Comparison of Video Shot Boundary Detection Techniques. *Journal of Electronic Imaging*, 5(2):122–128. SPIE - International Society for Optical Engineering.

Budíková, P., Botorek, J., Batko, M., and Zezula, P. (2014). DISA at ImageCLEF 2014 Revised: Search-based Image Annotation with DeCAF Features. *CoRR*, abs/1409.4627.

Budíkova, P., Batko, M., and Zezula, P. (2011). Evaluation Platform for Content-based Image Retrieval Systems. In *Research and Advanced Technology for Digital Libraries: International Conference on Theory and Practice of Digital Libraries, TPDL 2011*, pages 130–142, Berlin, Germany.

Budíková, P., Botorek, J., Batko, M., and Zezula, P. (2014). DISA at ImageCLEF 2014: The Search-based Solution for Scalable Image Annotation. In *CLEF 2014 Evaluation Labs and Workshop*, pages 360–371, Sheffield, UK.

Callan, J. P. (1994). Passage-level Evidence in Document Retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 302–310, Dublin, Ireland.

Campanella, M., Leonardi, R., and Migliorati, P. (2005). The Future-Viewer Visual Environment for Semantic Characterization of Video Sequences. In *Proceedings of the 2005 International Conference on Image Processing, ICIP 2005*, volume I, pages 1209–1212, Genova, Italy.

Candan, K. S. and Sapino, M. L. (2010). *Data Management for Multimedia Retrieval*, chapter Evaluation of Retrieval, pages 380–397. Cambridge University Press, New York, NY, USA.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., and Wellner, P. (2006). The AMI Meeting Corpus: A Pre-announcement. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, pages 28–39, Edinburgh, UK.

Chatfield, K., Arandjelović, R., Parkhi, O., and Zisserman, A. (2015). On-the-fly Learning for Visual Search of Large-scale Image and Video Datasets. *International Journal of Multimedia Information Retrieval*, 4:75–93. Springer London.

Chen, S., Curtis, K., Racca, D. N., Zhou, L., Jones, G. J. F., and O'Connor, N. E. (2015). DCU ADAPT @ TRECVid 2015: Video Hyperlinking Task. In *2015 TREC Video Retrieval Evaluation Notebook Papers and Slides*, Gaithersburg, MD, USA.

Chen, S., Jones, G. J. F., and O'Connor, N. E. (2013). DCU Linking Runs at MediaEval 2013: Search and Hyperlinking Task. In *MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain.

Chen, S., Jones, G. J. F., and O'Connor, N. E. (2014). DCU Linking Runs at MediaEval 2014: Search and Hyperlinking Task. In *MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain.

Cheng, Z., Li, X., Shen, J., and Hauptmann, A. G. (2015). CMU-SMU@TRECVID 2015: Video Hyperlinking. In *2015 TREC Video Retrieval Evaluation Notebook Papers and Slides*, Gaithersburg, MD, USA.

Chiu, J. and Rudnicky, A. (2014). LACS System Analysis on Retrieval Models for the MediaEval 2014 Search and Hyperlinking Task. In *MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain.

Choi, F. Y. Y. (2000). Advances in Domain Independent Linear Text Segmentation. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference, NAACL 2000*, pages 26–33, Seattle, WA, USA.

Chum, O., Philbin, J., Isard, M., and Zisserman, A. (2007). Scalable Near Identical Image and Shot Detection. In *Proceedings of the 6th ACM international conference on Image and video retrieval, CIVR '07*, pages 549–556, Amsterdam, The Netherlands.

Cieri, C., Strassel, S., Graff, D., Martey, N., Rennert, K., and Liberman, M. (2002). *Corpora for Topic Detection and Tracking*, pages 33–66. Springer US, Boston, MA, USA.

Ciresan, D., Meier, U., and Schmidhuber, J. (2012). Multi-column Deep Neural Networks for Image Classification. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3642–3649, Providence, RI, USA.

Cisco (2015). Cisco Visual Networking Index: Forecast and Methodology, 2014–2019. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.pdf.

Cleverdon, C. and Kean, M. (1968). Factors Determining the Performance of Indexing Systems. Technical report, Aslib Cranfield Research Project, Cranfield, England.

Cobârzan, C., Schoeffmann, K., Bailer, W., Hürst, W., Blažek, A., Lokoč, J., Vrochidis, S., Barthel, K. U., and Rossetto, L. (2017). Interactive Video Search Tools: A Detailed Analysis of the Video Browser Showdown 2015. *Multimedia Tools and Applications*, 76(4):5539–5571. Springer US.

Courtney, J. D. (1997). Automatic Video Indexing via Object Motion Analysis. *Pattern Recognition*, 30(4):607 – 625. Elsevier.

Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., and Sampath, D. (2010). The YouTube Video Recommendation System. In *Proceedings of the fourth ACM conference on Recommender systems, RecSys '10*, pages 293–296, Barcelona, Spain.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Li, F.-F. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009*, pages 248–255, Miami, FL, USA.

Dielmann, A. and Renals, S. (2005). Multistream Dynamic Bayesian Network for Meeting Segmentation. In *Machine Learning for Multimodal Interaction: First International Workshop, MLMI 2004, Revised Selected Papers*, pages 76–86, Martigny, Switzerland.

Domo (2013). Data Never Sleeps. https://www.domo.com/learn/infographic-data-never-sleeps.

Domo (2015). Data Never Sleeps 3.0. https://www.domo.com/learn/data-never-sleeps-3-0.

Eidenberger, H. (2012). *Handbook of Multimedia Information Retrieval*, chapter Fundamental Media Understanding, Introduction, pages 3–7. Atpress.

Eisenstein, J. and Barzilay, R. (2008). Bayesian Unsupervised Topic Segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 334–343, Honolulu, Hawaii.

Eskevich, M., Aly, R., Nicolás, D. R., Ordelman, R., Chen, S., and Jones, G. J. F. (2014a). Search and Hyperlinking 2014 Overview. http://www.slideshare.net/mariaeskevich/search-and-hyperlinking-me14sh-task-overviewmero.

Eskevich, M., Aly, R., Ordelman, R., Racca, D. N., Chen, S., and Jones, G. J. F. (2015). SAVA at MediaEval 2015: Search and Anchoring in Video Archives. In *MediaEval 2015 Multimedia Benchmark Workshop*, Wurzen, Germany.

Eskevich, M., Aly, R., Racca, D. N., Ordelman, R., Chen, S., and Jones, G. J. F. (2014b). The Search and Hyperlinking Task at MediaEval 2014. In *MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain.

Eskevich, M. and Huet, B. (2015). EURECOM @ SAVA2015: Visual Features for Multimedia Search. In *MediaEval 2015 Multimedia Benchmark Workshop*, Wurzen, Germany.

Eskevich, M. and Jones, G. J. F. (2014). Exploring Speech Retrieval from Meetings using the AMI Corpus. *Computer Speech & Language*, 28(5):1021–1044. Elsevier.

Eskevich, M., Jones, G. J. F., Aly, R., Ordelman, R., Chen, S., Nadeem, D., Guinaudeau, C., Gravier, G., Sébillot, P., De Nies, T., Debevere, P., Van de Walle, R., Galuščáková, P., Pecina, P., and Larson, M. (2013a). Multimedia Information Seeking through Search and Hyperlinking. In *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval*, pages 287–294, Dallas, TX, USA.

Eskevich, M., Jones, G. J. F., Chen, S., Aly, R., and Ordelman, R. (2013b). The Search and Hyperlinking Task at MediaEval 2013. In *MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain.

Eskevich, M., Jones, G. J. F., Chen, S., Aly, R., Ordelman, R., and Larson, M. (2012a). Search and Hyperlinking Task at MediaEval 2012. http://www.slideshare.net/MediaEval2012/search-and-hyperlinking-task-at-mediaeval-2012.

Eskevich, M., Jones, G. J. F., Chen, S., Aly, R., Ordelman, R., and Larson, M. (2012b). Search and Hyperlinking Task at MediaEval 2012. In *MediaEval 2012 Multimedia Benchmark Workshop*, Pisa, Italy.

Eskevich, M., Jones, G. J. F., Larson, M., and Ordelman, R. (2012c). Creating a Data Collection for Evaluating Rich Speech Retrieval. In *The Eighth international conference on Language Resources and Evaluation (LREC) 2012*, Istanbul, Turkey.

Eskevich, M., Jones, G. J. F., Wartena, C., Larson, M., Aly, R., Verschoor, T., and Ordelman, R. (2012d). Comparing Retrieval Effectiveness of Alternative Content Segmentation Methods for Internet Video Search. In *10th International Workshop on Content-Based Multimedia Indexing (CBMI), 2012*, Annecy, France.

Eskevich, M., Magdy, W., and Jones, G. J. F. (2012e). New Metrics for Meaningful Evaluation of Informally Structured Speech Retrieval. In *Advances in Information Retrieval: 34th European Conference on IR Research, ECIR 2012*, pages 170–181, Barcelona, Spain.

Eskevich, M., Nguyen, H., Sahuguet, M., and Huet, B. (2015). Hyper Video Browser: Search and Hyperlinking in Broadcast Media. In *MM 2015, ACM Multimedia Conference*, pages 817–818, Brisbane, Australia.

Eyben, F., Weninger, F., Gross, F., and Schuller, B. (2013). Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor. In *MM '13 Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838, Barcelona, Spain.

Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., and Yanker, P. (1995). Query by Image and Video Content: The QBIC System. *Computer*, 28(9):23–32. IEEE Computer Society Press.

Furht, B., editor (2008). *Encyclopedia of Multimedia*, chapter Video Near-duplicate Detection, pages 923–929. Springer US.

Galuščáková, P., Batko, M., Kruliš, M., Lokoč, J., Novák, D., and Pecina, P. (2015). CUNI at TRECVID 2015 Video Hyperlinking Task. In *2015 TREC Video Retrieval Evaluation Notebook Papers and Slides*, Gaithersburg, MD, USA.

Galuščáková, P., Batko, M., Čech, J., Matas, J., Novák, D., and Pecina, P. (2017). Visual Descriptors in Methods for Video Hyperlinking. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 294–300, Bucharest, Romania.

Galuščáková, P., Kruliš, M., Lokoč, J., and Pecina, P. (2014). CUNI at MediaEval 2014 Search and Hyperlinking Task: Visual and Prosodic Features in Hyperlinking. In *MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain.

Galuščáková, P. and Pecina, P. (2012). CUNI at MediaEval 2012 Search and Hyperlinking Task. In *MediaEval 2012 Multimedia Benchmark Workshop*, Pisa, Italy.

Galuščáková, P. and Pecina, P. (2013a). CUNI at MediaEval 2013 Search and Hyperlinking Task. In *MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain.

Galuščáková, P. and Pecina, P. (2013b). CUNI at MediaEval 2013 Similar Segments in Social Speech Task. In *MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain.

Galuščáková, P. and Pecina, P. (2014a). CUNI at MediaEval 2014 Search and Hyperlinking Task: Search Task Experiments. In *MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain.

Galuščáková, P. and Pecina, P. (2014b). Experiments with Segmentation Strategies for Passage Retrieval in Audio-Visual Documents. In *ICMR '14 Proceedings of International Conference on Multimedia Retrieval*, pages 217–224, Glasgow, UK.

Galuščáková, P. and Pecina, P. (2015). Audio Information for Hyperlinking of TV Content. In *SLAM '15 Proceedings of the Third Edition Workshop on Speech, Language & Audio in Multimedia*, pages 27–30, Brisbane, Australia.

Galuščáková, P., Saleh, S., and Pecina, P. (2016). SHAMUS: UFAL Search and Hyperlinking Multimedia System. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016*, pages 853–856, Padua, Italy.

Galuščáková, P. and Pecina, P. (2015). CUNI at MediaEval 2015 Search and Anchoring in Video Archives: Anchoring via Information Retrieval. In *MediaEval 2015 Multimedia Benchmark Workshop*, Wurzen, Germany.

Galuščáková, P., Pecina, P., and Hajič, J. (2012). Penalty Functions for Evaluation Measures of Unsegmented Speech Retrieval. In *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics: Third International Conference of the CLEF Initiative, CLEF 2012*, volume 7488 of *LNCS*, pages 100–111, Rome, Italy.

Galuščáková, P., Saleh, S., and Pecina, P. (2016). UFAL Search and Hyperlinking Multimedia System. https://ufal.mff.cuni.cz/shamus.

García, F., Sanchis, E., Calvo, M., Pla, F., and Hurtado, L.-F. (2013). ELiRF at MediaEval 2013: Similar Segments in Social Speech Task. In *MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain.

Ghias, A., Logan, J., Chamberlin, D., and Smith, B. C. (1995). Query by Humming: Musical Information Retrieval in an Audio Database. In *Proceedings of the third ACM international conference on Multimedia MULTIMEDIA '95*, pages 231–236, San Francisco, CA, USA.

Gibbs, A. J. and Mcintyre, G. A. (1970). The Diagram, a Method for Comparing Sequences. *European Journal of Biochemistry*, 16(1):1–11. Blackwell Publishing Ltd.

Goëau, H., Thièvre, J., Viaud, M.-L., and Pellerin, D. (2007). Interactive Visualization Tool with Graphic Table of Video Contents. In *2007 IEEE International Conference on Multimedia and Expo*, pages 807–810, Beijing, China.

Gomez-Uribe, C. A. and Hunt, N. (2015). The Netflix Recommender System: Algorithms, Business Value, and Innovation. *ACM Transactions on Management Information Systems*, 6(4):13:1–13:19. ACM.

Gordon, A. S. and Domeshek, E. A. (2001). Retrieval Interfaces for Video Databases. In *Proceedings of the AAAI fall Symposium on AI applications in knowledge Navigation and retrieval*, pages 45–51, Cambridge, MA, USA.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18. ACM.

Hauptmann, A. G. and Witbrock, M. J. (1998). Story Segmentation and Detection of Commercials in Broadcast News Video. In *Proceedings. IEEE International Forum on Research and Technology Advances in Digital Libraries, 1998*, pages 168–179, Santa Barbara, CA, USA.

Hearst, M. A. (1997). TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistics*, 23(1):33–64. MIT Press.

Heesch, D., Howarth, P., Magalhaes, J., May, A., Pickering, M., Yavlinsky, A., and Rüger, S. (2004). Video Retrieval using Search and Browsing. In *The Thirteenth Text Retrieval Conference (TREC 2004)*, pages 15–19, Gaithersburg, MD, USA.

Hetland, M. L., Skopal, T., Lokoč, J., and Beecks, C. (2013). Ptolemaic Access Methods: Challenging the Reign of the Metric Space Model. *Information Systems*, 38(7):989–1006. Elsevier.

Hiemstra, D. (2001). *Using Language Models for Information Retrieval*. PhD thesis, University of Twente, Enschede, Netherlands.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition. *Signal Processing Magazine*, 29(6):82 – 97. IEEE.

Hirschberg, J. and Litman, D. (1993). Empirical Studies on the Disambiguation of Cue Phrases. *Computational Linguistics*, 19(3):501–530. MIT Press.

Hsueh, P.-Y. and Moore, J. D. (2007). Combining Multiple Knowledge Sources for Dialogue Segmentation in Multimedia Archives. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 1016–1023, Prague, Czech Republic.

Iyengar, G., Duygulu, P., Feng, S., Ircing, P., Khudanpur, S. P., Klakow, D., Krause, M. R., Manmatha, R., Nock, H. J., Petkova, D., Pytlik, B., and Virga, P. (2005). Joint Visual-text Modeling for Automatic Retrieval of Multimedia Documents. In *Proceedings of the 13th annual ACM international conference on Multimedia MULTIMEDIA '05*, pages 21–30, Singapore.

James, S. and Collomosse, J. (2014). Interactive Video Asset Retrieval Using Sketched Queries. In *Proceedings of the 11th European Conference on Visual Media Production CVMP '14*, pages 11:1–11:8, London, UK.

Jeong, M. and Titov, I. (2010). Multi-document Topic Segmentation. In *Proceedings of the 19th ACM international conference on Information and knowledge management CIKM '10*, pages 1119–1128, Toronto, Canada.

Kaszkiel, M. and Zobel, J. (1997). Passage Retrieval Revisited. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR '97*, pages 178–185, Philadelphia, PA, USA.

Kaszkiel, M. and Zobel, J. (2001). Effective Ranking with Arbitrary Passages. *Journal of the American Society for Information, Science and Technology*, 52(4):344–364. Wiley-Blackwell.

Kato, T., Kurita, T., Otsu, N., and Hirata, K. (1992). A Sketch Retrieval Method for Full Color Image Database-query by Visual Example. In *11th IAPR International Conference on Pattern Recognition, Vol.I. Conference A: Computer Vision and Applications*, pages 530–533, The Hague, Netherlands.

Kauchak, D. and Chen, F. (2005). Feature-based Segmentation of Narrative Documents. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing FeatureEng '05*, pages 32–39, Ann Arbor, MI, USA.

Kelm, P., Schmiedeke, S., and Sikora, T. (2009). Feature-based Video Key Frame Extraction for Low Quality Video Sequences. In *10th Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS '09*, pages 25–28, London, UK.

Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J. (1998). On Combining Classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239. IEEE Computer Society.

Kozima, H. (1993). Text Segmentation Based On Similarity Between Words. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics ACL '93*, pages 286–288, Columbus, Ohio.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105, Lake Tahoe, NV, USA.

Kruliš, M., Lokoč, J., and Skopal, T. (2013). Efficient Extraction of Feature Signatures Using Multi-GPU Architecture. In *Advances in Multimedia Modeling: 19th International Conference, MMM 2013, Part II*, volume 7733 of *LNCS*, pages 446–456, Huangshan, China.

Kruliš, M., Skopal, T., Lokoč, J., and Beecks, C. (2012). Combining CPU and GPU Architectures for Fast Similarity Search. *Distributed and Parallel Databases*, 30(3-4):179–207. Springer US.

Kruliš, M., Lokoč, J., and Skopal, T. (2016). Efficient Extraction of Clustering-based Feature Signatures Using GPU Architectures. *Multimedia Tools and Applications*, 75(13):8071–8103. Springer US.

Kuboň, D., Blažek, A., Lokoč, J., and Skopal, T. (2016). Multi-sketch Semantic Video Browser. In Tian, Q., Sebe, N., Qi, G.-J., Huet, B., Hong, R., and Liu, X., editors, *MultiMedia Modeling: 22nd International Conference, MMM 2016, Part II*, pages 406–411, Miami, FL, USA.

Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K. (2009). Attribute and Simile Classifiers for Face Verification. In *2009 IEEE 12th International Conference on Computer Vision*, pages 365–372, Kyoto, Japan.

Lalmas, M. (2011). Introduction to Information Retrieval. Summer School Lecture. http://www.dcs.gla.ac.uk/ mounia/mounia.pdf.

Lamel, L. (2012). Multilingual Speech Processing Activities in Quaero: Application to Multimedia Search in Unstructured Data. In *The Fifth International Conference Human Language Technologies – The Baltic Perspective*, pages 1–8, Tartu, Estonia.

Lamel, L. and Gauvain, J.-L. (2008). Speech Processing for Audio Indexing. In *Advances in Natural Language Processing: 6th International Conference, GoTAL 2008*, pages 4–15, Gothenburg, Sweden.

Lanchantin, P., Bell, P.-J., Gales, M.-J.-F., Hain, T., Liu, X., Long, Y., Quinnell, J., Renals, S., Saz, O., Seigel, M.-S., Swietojanski, P., and Woodland, P.-C. (2013). Automatic Transcription of Multi-genre Media Archives. In *Proceedings of the First Workshop on Speech, Language and Audio in Multimedia*, pages 26–31, Marseille, France.

Larson, M., Eskevich, M., Ordelman, R., Kofler, C., Schmiedeke, S., and Jones, G. J. (2011). Overview of MediaEval 2011 Rich Speech Retrieval Task and Genre Tagging Task. In *MediaEval 2011 Multimedia Benchmark Workshop*, Pisa, Italy.

Larson, M., Ionescu, B., Sjöberg, M., Anguera, X., Poignant, J., Riegler, M., Eskevich, M., Hauff, C., Sutcliffe, R., Jones, G. J. F., Yang, Y.-H., Soleymani, M., and Papadopoulos, S., editors (2015). *MediaEval 2015 Multimedia Benchmark Workshop*, Wurzen, Germany.

Larson, M. A. and Jones, G. J. (2012a). *Spoken Content Retrieval: A Survey of Techniques and Technologies*, volume 5 of *Foundations and Trends in Information Retrieval*. Now Publishers Inc.

Larson, M. A. and Jones, G. J. (2012b). *Spoken Content Retrieval: A Survey of Techniques and Technologies*, chapter Challenges for SCR, pages 273–276. Volume 5 of (Larson and Jones, 2012a).

Le, H., Q.M. Bui and, B. H., Červenková, B., Bouchner, J., Apostolidis, E., Markatopoulou, F., Pournaras, A., Mezaris, V., Stein, D., Eickeler, S., and Stadtschnitzer, M. (2014). LinkedTV at MediaEval 2014 Search and Hyperlinking Task. In *MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain.

Levene, M. (2010). *An Introduction to Search Engines and Web Navigation*. Wiley Publishing, 2nd edition.

Levow, G.-A. (2013). UWCL at MediaEval 2013: Similar Segments in Social Speech Task. In *MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain.

Li, H., Jou, B., Ellis, J. G., Morozoff, D., and Chang, S.-F. (2013). News Rover: Exploring Topical Structures and Serendipity in Heterogeneous Multimedia News. In *MM '13 Proceedings of the 21st ACM international conference on Multimedia*, pages 449–450, Barcelona, Spain.

Lienhart, R. (1999). Comparison of Automatic Shot Boundary Detection Algorithms. In *Storage and Retrieval for Image and Video Databases*, pages 290–301, San Jose, CA, USA.

Lin, J. J., Quan, D., Sinha, V., Bakshi, K., Huynh, D., Katz, B., and Karger, D. R. (2003). What Makes a Good Answer? The Role of Context in Question Answering. In *Proceedings of INTERACT 2003*, pages 25–32, Zurich, Switzerland.

Liu, B. and Oard, D. W. (2006). One-sided Measures for Evaluating Ranked Retrieval Effectiveness with Spontaneous Sonversational Speech. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR '06*, pages 673–674, Seattle, WA, USA.

Liu, X. and Croft, W. B. (2002). Passage Retrieval Based On Language Models. In *Proceedings of the eleventh international conference on Information and knowledge management CIKM '02*, pages 375–382, McLean, VA, USA.

Lokaj, M., Stiegler, H., and Bailer, W. (2013). TOSCA-MP at Search and Hyperlinking of Television Content Task. In *MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain.

Lokoč, J., Schoeffmann, K., and del Fabro, M. (2015). Dynamic Hierarchical Visualization of Keyframes in Endoscopic Video. In *MultiMedia Modeling: 21st International Conference, MMM 2015, Part II*, pages 291–294, Reykjavik, Iceland.

Malioutov, I. and Barzilay, R. (2006). Minimum Cut Model for Spoken Lecture Segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 25–32, Sydney, Australia.

Malioutov, I., Park, A., Barzilay, R., and Glass, J. R. (2007). Making Sense of Sound: Unsupervised Topic Segmentation over Acoustic Input. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 504–511, Prague, Czech Republic.

Manning, C. D. (1998). Rethinking Text Segmentation Models: An Information Extraction Case Study. Technical report, University of Sydney, Sydney, Australia.

Manning, C. D., Raghavan, P., and Schütze, H. (2008a). *Introduction to Information Retrieval*. Cambridge University Press.

Manning, C. D., Raghavan, P., and Schütze, H. (2008b). *Introduction to Information Retrieval*, chapter Evaluation of ranked retrieval results, pages 151–175. In (Manning et al., 2008a).

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330. MIT Press.

Melucci, M. (1998). Passage Retrieval: A Probabilistic Technique. *Information Processing and Management*, 34(1):43–68. Pergamon Press, Inc.

Merriam-Webster Online (2009). Merriam-Webster Online Dictionary. http://www.merriam-webster.com.

Mihalcea, R. and Csomai, A. (2007). Wikify!: Linking Documents to Encyclopedic Knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management CIKM '07*, pages 233–242, Lisbon, Portugal.

Mikolov, T., tau Yih, S. W., and Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, GA, USA.

Miller, G. (2014). Data From a Century of Cinema Reveals How Movies Have Evolved. http://www.wired.com/2014/09/cinema-is-evolving.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41. ACM.

Misra, H., Yvon, F., Jose, J. M., and Cappe, O. (2009). Text Segmentation via Topic Modeling: An Analytical Study. In *Proceedings of the 18th ACM conference on Information and knowledge management CIKM '09*, pages 1553–1556, Hong Kong, China.

Mittendorf, E. and Schäuble, P. (1994). Document and Passage Retrieval Based on Hidden Markov Models. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR '94*, pages 318–327, Dublin, Ireland.

Mohri, M., Moreno, P., and Weinstein, E. (2010). Discriminative Topic Segmentation of Text and Speech. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 533–540, Chia Laguna Resort, Sardinia, Italy.

Morris, J. and Hirst, G. (1988). Lexical Cohesion, the Thesaurus, and the Structure of Text. Technical report, Computer Systems Research Institute, University of Toronto, Toronto, Canada.

Moumtzidou, A., Avgerinakis, K., Apostolidis, E., Markatopoulou, F., Apostolidis, K., Mironidis, T., Vrochidis, S., Mezaris, V., Kompatsiaris, I., and Patras, I. (2015). VERGE: A Multimodal Interactive Video Search Engine. In *MultiMedia Modeling: 21st International Conference, MMM 2015, Part II*, pages 249–254, Sydney, Australia.

Moyal, A., Aharonson, V., Tetariy, E., and Gishri, M. (2013). *Phonetic Search Methods for Large Speech Databases*, chapter Keyword Spotting Out of Continuous Speech, pages 7–11. Springer New York.

National Institute of Standards and Technology (2006). The Spoken Term Detection (STD) 2006 Evaluation Plan. Technical report, Gaithersburg, MD, USA.

Nguyen, V. C., Nguyen, L. M., and Shimazu, A. (2011). Improving Text Segmentation with Non-systematic Semantic Relation. In *Computational Linguistics and Intelligent Text Processing: 12th International Conference, CICLing 2011, Part I*, pages 304–315, Tokyo, Japan.

Niaz, U., Merialdo, B., Tanase, C., Eskevich, M., and Huet, B. (2015). EURECOM at TrecVid 2015: Semantic Indexing and Video Hyperlinking Tasks. In *Proceedings of TRECVID 2015*, Gaithersburg, MD, USA.

Nies, T. D., Neve, W. D., Mannens, E., and de Walle, R. V. (2013). Ghent University-iMinds at MediaEval 2013: An Unsupervised Named Entity-based Similarity Measure for Search and Hyperlinking. In *MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain.

Novák, D., Batko, M., and Zezula, P. (2015). Large-scale Image Retrieval using Neural Net Descriptors. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR '15*, Santiago, Chile.

Ordelman, R., Aly, R., Eskevich, M., Huet, B., Jones, G. J. F., and Racca, D. N. (2015a). Video Hyperlinking TRECVid 2015.

Ordelman, R., Eskevich, M., Aly, R., Huet, B., and Jones, G. J. F. (2015b). Defining and Evaluating Video Hyperlinking for Navigating Multimedia Archives. In *Companion Proceedings of the 24th International Conference on World Wide Web, WWW '15*, pages 727–732.

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., and Lioma, C. (2006). Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proc. of SIGIR Workshop on Open Source Information Retrieval*, pages 18–25, Seattle, WA, USA.

Over, P., Awad, G., Michel, M., Fiscus, J., Kraaij, W., Smeaton, A. F., Quéenot, G., and Ordelman, R. (2015). TRECVID 2015 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *2015 TREC Video Retrieval Evaluation Notebook Papers and Slides*, Gaithersburg, MD, USA.

Pang, L. and Ngo, C.-W. (2015). VIREO @ TRECVID 2015: Video Hyperlinking (LNK). In *2015 TREC Video Retrieval Evaluation Notebook Papers and Slides*, Gaithersburg, MD, USA.

Papka, R. and Allen, J. (1997). Why Bigger Windows Are Better Than Smaller Ones. Technical report, University of Massachusetts, Amherst, MA, USA.

Pappas, N. and Popescu-Belis, A. (2013). Combining Content with User Preferences for TED Lecture Recommendation. In *11th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 47–52, Veszprém, Hungary.

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep Face Recognition. In *Proceedings of the British Machine Vision Conference 2015*, pages 41.1–41.12, Swansea, UK.

Paróczi, Z., Fodor, B., and Szücs, G. (2014). DCLab at MediaEval2014 Search and Hyperlinking Task. In *MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain.

Patel, B. V. and Meshram, B. B. (2012). Content Based Video Retrieval Systems. *CoRR*, abs/1205.1641.

Pevzner, L. and Hearst, M. A. (2002). A Critique and Improvement of an Evaluation Metric for Text Segmentation. *Computational Linguistics*, 28(1):19–36. MIT Press.

Ponte, J. M. and Croft, W. B. (1997). Text Segmentation by Topic. In *Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries ECDL '97*, volume 1324 of *LNCS*, pages 113–125, Pisa, Italy.

Porter, M. F. (1997). *Readings in Information Retrieval*, chapter An Algorithm for Suffix Stripping, pages 313–316. Morgan Kaufmann Publishers Inc.

Potapov, D., Douze, M., Harchaoui, Z., and Schmid, C. (2014). Category-Specific Video Summarization. In *Computer Vision – ECCV 2014: 13th European Conference, Part VI*, volume 8694 of *LNCS*, Zurich, Switzerland.

Preston, J., Hare, J., Samangooei, S., Davies, J., Jain, N., and Dupplaw, D. (2013). A Unified, Modular and Multimodal Approach to Search and Hyperlinking Video. In *MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain.

Pye, D., Hollinghurst, N. J., Mills, T. J., and Wood, K. R. (1998). Audio-Visual Segmentation for Content-Based Retrieval. In *5th International Conference on Spoken Language Processing*, Sydney, Australia.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Racca, D. N., Eskevich, M., and Jones, G. J. (2014). DCU Search Runs at MediaEval 2014 Search and Hyperlinking. In *MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain.

Racca, D. N. and Jones, G. J. F. (2015). Evaluating Search and Hyperlinking: An Example of the Design, Test, Refine Cycle for Metric Development. In *MediaEval 2015 Multimedia Benchmark Workshop*, Wurzen, Germany.

Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G. R., Levy, R., and Vasconcelos, N. (2010). A New Approach to Cross-modal Multimedia Retrieval. In *Proceedings of the 18th ACM international conference on Multimedia MM '10*, pages 251–260, Firenze, Italy.

Revaud, J., Douze, M., Schmid, C., and Jégou, H. (2013). Event Retrieval in Large Video Collections with Circulant Temporal Encoding. In *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2459–2466, Portland, OR, USA.

Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B. (2010). *Recommender Systems Handbook*, chapter Introduction to Recommender Systems Handbook, pages 1–35. Springer-Verlag New York, Inc.

Roberts, I. and Gaizauskas, R. J. (2004). Evaluating Passage Retrieval Approaches for Question Answering. In *Advances in Information Retrieval: 26th European Conference on IR Research, ECIR 2004*, volume 2997 of *LNCS*, pages 72–84, Sunderland, UK.

Robertson, S. E. and Walker, S. (1994). Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '94*, pages 232–241, Dublin, Ireland.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M., and Gatford, M. (1994). Okapi at TREC-3. In *Proceedings of the 3rd Text REtrieval Conference*, pages 109–126, Gaithersburg, MD, USA.

Rossetto, L., Giangreco, I., Heller, S., Tanase, C., Schuldt, H., Dupont, S., Seddati, O., Sezgin, T. M., Altiok, O. C., and Sahillioglu, Y. (2016). IMOTION - Searching for Video Sequences Using Multi-Shot Sketch Queries. In *MultiMedia Modeling: 22nd International Conference, MMM 2016, Part II*, volume 9517 of *LNCS*, pages 377–382, Miami, FL, USA.

Rousseau, A., Bougares, F., Deléglise, P., Schwenk, H., and Estève, Y. (2011). LIUM's systems for the IWSLT 2011 Speech Translation Tasks. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 79–85, San Francisco, CA, USA.

Rousseau, A., Deléglise, P., and Estève, Y. (2014). Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling And More TED Talks. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3935–3939, Reykjavik, Iceland.

Rubner, Y. and Tomasi, C. (2001). *Perceptual Metrics for Image Database Navigation*, chapter Color Signatures, pages 32–33. Kluwer Academic Publishers.

Rüger, S. (2010). *Multimedia Information Retrieval*, chapter Basic Multimedia Search Technologies, pages 13–40. Synthesis Lectures on Information Concepts, Retrieval and Services. Morgan & Claypool Publishers.

Sahuguet, M., Huet, B., Červenková, B., Apostolidis, E., Mezaris, V., Stein, D., Eickeler, S., José, Garcia, L. R., Troncy, R., and Pikora, L. (2013). LinkedTV at MediaEval 2013 Search and Hyperlinking Task. In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain.

Salton, G., Allan, J., and Buckley, C. (1993). Approaches to Passage Retrieval in Full Text Information Systems. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '93*, pages 49–58, Pittsburgh, PA, USA.

Schein, A. I., Popescul, A., Ungar, L. H., and Pennock, D. M. (2002). Methods and Metrics for Cold-start Recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '02*, pages 253–260, Tampere, Finland.

Schmiedeke, S., Xu, P., Ferrané, I., Eskevich, M., Kofler, C., Larson, M. A., Estève, Y., Lamel, L., Jones, G. J. F., and Sikora, T. (2013). Blip10000: A Social Video Dataset containing SPUG Content for Tagging and Retrieval. In *Proceedings of the 4th ACM Multimedia Systems Conference, MMSys '13*, pages 96–101, Oslo, Norway.

Schoeffmann, K. (2014). A User-Centric Media Retrieval Competition: The Video Browser Showdown 2012-2014. *IEEE MultiMedia*, 21(4):8–13. IEEE.

Schoeffmann, K., Hopfgartner, F., Marques, O., Boeszoermenyi, L., and Jose, J. M. (2010a). Video Browsing Interfaces and Applications: A Review. *SPIE Reviews*, 1(1). SPIE.

Schoeffmann, K., Taschwer, M., and Boeszoermenyi, L. (2010b). The Video Explorer – A Tool for Navigation and Searching within a Single Video based on Fast Content Analysis. In *Proceedings of the first annual ACM SIGMM conference on Multimedia systems, MMSys '10*, pages 247–258, Phoenix, AZ, USA.

Shapiro, L. G. and Stockman, G. C. (2001). *Computer Vision*, chapter Content-Based Image Retrieval, pages 249–274. Prentice Hall.

Simon, A., Guinaudeau, C., Sébillot, P., and Gravier, G. (2014). Investigating Domain-independent NLP Techniques for Precise Target Selection in Video Hyperlinking. In *2nd International Workshop on Speech, Language and Audio in Multimedia, SLAM*, pages 19–23, Penang, Malaysia.

Simon, A.-R., Bois, R., Gravier, G., Sébillot, P., Morin, E., and Moens, S. (2015a). Hierarchical Topic Models for Language-based Video Hyperlinking. In *SLAM '15 Proceedings of the Third Edition Workshop on Speech, Language & Audio in Multimedia*, pages 27–30, Brisbane, Australia.

Simon, A.-R. and Gravier, G. (2015). IRISA at MediaEval 2015: Search and Anchoring in Video Archives Task. In *MediaEval 2015 Multimedia Benchmark Workshop*, Wurzen, Germany.

Simon, A.-R., Sicre, R., Bois, R., Gravier, G., Sébillot, P., and Morin, E. (2015b). IRISA at TrecVid2015: Leveraging Multimodal LDA for Video Hyperlinking. In *Proceedings of TRECVID 2015*, Gaithersburg, MD, USA.

Smeaton, A., Kraaij, W., and Over, P. (2003). TRECVID 2003 - An Overview. In *TRECVID 2003 - Text REtrieval Conference TRECVID Workshop*, Gaithersburg, MD, USA.

Smeaton, A. F., Kraaij, W., and Over, P. (2004). TRECVID 2004 - An Overview. In *TRECVID 2004 - Text REtrieval Conference TRECVID Workshop*, Gaithersburg, MD, USA.

Snoek, C. G., Worring, M., de Rooij, O., van de Sande, K. E., Yan, R., and Hauptmann, A. G. (2008). VideOlympics: Real-Time Evaluation of Multimedia Retrieval Systems. *IEEE MultiMedia*, 15(1):86–91. IEEE.

Snoek, C. G., Worring, M., Koelma, D. C., and Smeulders, A. W. (2007). A Learned Lexicon-Driven Paradigm for Interactive Video Retrieval. *IEEE Transactions on Multimedia*, 9(2):280–292. IEEE.

Snoek, C. G. M. and Worring, M. (2009). Concept-Based Video Retrieval. *Foundations and Trends in Information Retrieval*, 2(4):215–322. Now Publishers Inc.

Spangler, T. (2015). Netflix Bandwidth Usage Climbs to Nearly 37% of Internet Traffic at Peak Hours. http://variety.com/2015/digital/news/netflix-bandwidth-usage-internet-traffic-1201507187.

Spousta, M. (2013). Featurama – A Library That Implements Various Sequence-labeling Algorithms. http://sourceforge.net/projects/featurama.

Stokes, N., Carthy, J., and Smeaton, A. F. (2004). SeLeCT: a Lexical Cohesion Based News Story Segmentation System. *AI Communications*, 17(1):3–12. IOS Press.

Sysomos Inc. (2010). Inside YouTube Videos. http://www.sysomos.com/reports/youtube.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going Deeper with Convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, Boston, MA, USA.

Tan, S., Ngo, C.-W., Tan, H.-K., and Pang, L. (2011). Cross Media Hyperlinking for Search Topic Browsing. In *Proceedings of the 19th ACM international conference on Multimedia, MM '11*, pages 243–252, Scottsdale, AZ, USA.

Tellex, S., Katz, B., Lin, J., Fernandes, A., and Marton, G. (2003). Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, SIGIR '03*, pages 41–47, Toronto, Canada.

Tiedemann, J. and Mur, J. (2008). Simple is Best: Experiments with Different Document Segmentation Strategies for Passage Retrieval. In *Proceedings of the 2nd workshop on Information Retrieval for Question Answering (IR4QA)*, pages 17–25, Manchester, UK.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 142–147, Edmonton, Canada.

Tommasi, T., Aly, R. B. N., McGuinness, K., Chatfield, K., Arandjelovic, R., Parkhi, O., Ordelman, R. J. F., Zisserman, A., and Tuytelaars, T. (2014). Beyond Metadata: Searching Your Archive Based on Its Audio-visual Content. In *Proceedings of the 2014 International Broadcasting Convention, IBC 2014*, Stevenage, UK.

Tür, G., Stolcke, A., Hakkani-Tür, D., and Shriberg, E. (2001). Integrating Prosodic and Lexical Cues for Automatic Topic Segmentation. *Computational Linguistics*, 27(1):31–57. MIT Press.

Vliegendhart, R., Liem, C. C. S., and Larson, M. (2015). Exploring Microblog Activity for the Prediction of Hyperlink Anchors in Television Broadcasts. In *MediaEval 2015 Multimedia Benchmark Workshop*, Wurzen, Germany.

Voorhees, E. (1999). The TREC-8 Question Answering Track Report. Technical report, National Institute of Standards and Technology, Gaithersburg, MD, USA.

Walgrove, A. (2015). The Explosive Growth of Online Video, in 5 Charts. https://contently.com/strategist/2015/07/06/the-explosive-growth-of-online-video-in-5-charts.

Wang, A. (2003). An Industrial-Strength Audio Search Algorithm. In *Proceedings of the 4 th International Conference on Music Information Retrieval*, Baltimore, MD, USA.

Wangphanitkun, K. (2015). Video Marketing Statistics & Trends 2015. http://syndacast.com/wp-content/uploads/2014/07/video-marketing-trends-2015.png.

Ward, N. G. and Werner, S. D. (2013). Data Collection for the Similar Segments in Social Speech Task. Technical report, University of Texas at El Paso, El Paso, TX, USA.

Ward, N. G., Werner, S. D., Novick, D. G., Shriberg, E. E., Oertel, C., Morency, L.-P., and Kawahara, T. (2013). The Similar Segments in Social Speech Task. In *MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain.

Wartena, C. (2012). Comparing Segmentation Strategies for Efficient Video Passage Retrieva. In *10th International Workshop on Content-Based Multimedia Indexing (CBMI), 2012*, pages 1–6, Annecy, France.

Wasserman, T. (2014). YouTube at 10: Under Siege but Still Dominant. http://mashable.com/2015/02/14/youtube-at-10-under-siege-but-still-dominant.

Werner, S. D. and Ward, N. G. (2013). Evaluating Prosody-Based Similarity Models for Information Retrieval. In *MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain.

Westerveld, T., de Vries, A. P., van Ballegooij, A., de Jong, F., and Hiemstra, D. (2003). A Probabilistic Multimedia Retrieval Model and Its Evaluation. *EURASIP Journal on Applied Signal Processing*, 2003:186–198. Hindawi Publishing Corp.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83. International Biometric Society.

Wilkins, P. (2009). *An Investigation Into Weighted Data Fusion for Content-Based Multimedia Information Retrieval*. PhD thesis, Dublin City University, Dublin, Ireland.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR*, abs/1609.08144.

Xiang, L. (2011). Hulu's Recommendation System. http://tech.hulu.com/blog/2011/09/19/recommendation-system.

Xiang, L. (2015). 500 Hours of Video Uploaded To YouTube Every Minute [Forecast]. http://www.reelseo.com/hours-minute-uploaded-youtube.

Xiao, C., Wang, C., Zhang, L., and Zhang, L. (2015). Sketch-based Image Retrieval via Shape Words. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, ICMR '15*, pages 571–574, Shanghai, China.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., and Zweig, G. (2016). Achieving Human Parity in Conversational Speech Recognition. *CoRR*, abs/1610.05256.

Xu, B., Liao, W., Liu, Z., Bao, W., Li, Y., Yang, D., Wang, S., Liu, H., Xia, Y., Wang, Y., and Chen, Z. (2015). IIPWHU@TRECVID 2015. In *2015 TREC Video Retrieval Evaluation Notebook Papers and Slides*, Gaithersburg, MD, USA.

Xu, J. and Croft, W. B. (1996). Query Expansion Using Local and Global Document Analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '96*, pages 4–11, Zurich, Switzerland.

Zabih, R., Miller, J., and Mai, K. (1995). A Feature-based Algorithm for Detecting and Classifying Scene Breaks. In *Proceedings of the third ACM international conference on Multimedia, MULTIMEDIA '95*, pages 189–200, San Francisco, CA, USA.

Publications

Maria Eskevich, Gareth J. F. Jones, Robin Aly, Roeland Ordelman, Shu Chen, Danish Nadeem, Camille Guinaudeau, Guillaume Gravier, Pascale Sébillot, Tom De Nies, Pedro Debevere, Rik Van de Walle, Petra Galuščáková, Pavel Pecina, and Martha Larson. Multimedia Information Seeking through Search and Hyperlinking. In *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval*, pages 287–294, Dallas, TX, USA, 2013.

Petra Galuščáková and Pavel Pecina. CUNI at MediaEval 2012 Search and Hyperlinking Task. In *MediaEval 2012 Multimedia Benchmark Workshop*, Pisa, Italy, 2012.

Petra Galuščáková and Pavel Pecina. CUNI at MediaEval 2013 Search and Hyperlinking Task. In *MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain, 2013a.

Petra Galuščáková and Pavel Pecina. CUNI at MediaEval 2013 Similar Segments in Social Speech Task. In *MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain, 2013b.

Petra Galuščáková and Pavel Pecina. Experiments with Segmentation Strategies for Passage Retrieval in Audio-Visual Documents. In *ICMR '14 Proceedings of International Conference on Multimedia Retrieval*, pages 217–224, Glasgow, UK, 2014a.

Petra Galuščáková and Pavel Pecina. CUNI at MediaEval 2014 Search and Hyperlinking Task: Search Task Experiments. In *MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain, 2014b.

Petra Galuščáková and Pavel Pecina. Audio Information for Hyperlinking of TV Content. In *SLAM '15 Proceedings of the Third Edition Workshop on Speech, Language & Audio in Multimedia*, pages 27–30, Brisbane, Australia, 2015.

PUBLICATIONS

Petra Galuščáková, Martin Kruliš, Jakub Lokoč, and Pavel Pecina. CUNI at MediaEval 2014 Search and Hyperlinking Task: Visual and Prosodic Features in Hyperlinking. In *MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain, 2014.

Petra Galuščáková, Michal Batko, Martin Kruliš, Jakub Lokoč, David Novák, and Pavel Pecina. CUNI at TRECVID 2015 Video Hyperlinking Task. In *2015 TREC Video Retrieval Evaluation Notebook Papers and Slides*, Gaithersburg, MD, USA, 2015.

Petra Galuščáková, Shadi Saleh, and Pavel Pecina. SHAMUS: UFAL Search and Hyperlinking Multimedia System. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016*, pages 853–856, Padua, Italy, 2016.

Petra Galuščáková, Michal Batko, Jan Čech, Jiří Matas, David Novák, and Pavel Pecina. Visual Descriptors in Methods for Video Hyperlinking. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 294–300, Bucharest, Romania, 2017.

Petra Galuščáková and Pavel Pecina. CUNI at MediaEval 2015 Search and Anchoring in Video Archives: Anchoring via Information Retrieval. In *MediaEval 2015 Multimedia Benchmark Workshop*, Wurzen, Germany, 2015.

Petra Galuščáková, Pavel Pecina, and Jan Hajič. Penalty Functions for Evaluation Measures of Unsegmented Speech Retrieval. In *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics: Third International Conference of the CLEF Initiative, CLEF 2012*, volume 7488 of *LNCS*, pages 100–111, Rome, Italy, 2012.