PH.D THESIS

INVESTIGATING MULTI-MODAL FEATURES IN THE DESIGN OF A MULTI-MEDIA HYPERLINKING FRAMEWORK

by

Shu Chen

School of Electronic Engineering

Supervisors:

Prof. Noel E. O'Connor, Prof. Gareth J.F. Jones

July 14th, 2016

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Ph.D in Electronic Engineering is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: _____ ID No.: _____
Shu Chen

Date: _____

Acknowledgments

There are a number of people without whom this thesis might not haven been published. I would like to express my sincerely thanks to these people.

First of all, I would like express my gratitude to my two supervisors, Prof. Noel E. O'Connor and Prof. Dr. Gareth J. F. Jones, for their enlightening suggestions and insightful encouragement which inspire me to widen my ideas and skills in academic from various perspectives. I appreciate their vast knowledge in Multimedia Processing, Information Retrieval and Natural Language Processing, which are not only the primary topic of my research but also the foundation of my future career.

I express my sincere thanks to European Commission's Seventh Framework Programme (FP7) as part of the AXES project (ICT-269980) to fund my Ph.D study.

My sincere thanks also goes to Dr.Kevin McGuinness, Dr. Maria Eskevich, and Dr. Dian Zhang for their stimulating discussions and generously support in my experimental investigation.

A very special thanks goes to my dear friends, Dr. Jie Yang, Dr. Shasha Li, Lei Zhang, Luo Luo, Ting Bi. We together came to Ireland from Wuhan University and shared more than 6 years on this land. I must appreciate your kind help in my study and life. The time being with you is my precious memory in my entire life.

Last but not the least, I would like to express my thanks and love to my family, my parents and my brother, for supporting me spiritually throughout writing this thesis and my life in general.

Abstract

Search, as a well-known information retrieval strategy, is widely researched and developed for academic and commercial usage. However, in the context of increasing amounts of multimedia data, search alone cannot satisfy user requirements for exploring multimedia resources. Therefore, preprocessing of multimedia resources is necessary to define potentially related documents to reduce retrieval time and improve the browsing efficiency. Using hyperlinks to connect relevant resources is widely used for multimedia collection. However, the definition of hyperlinks is usually based on textual information. For example, hyperlinks in Wikipedia link a term to relevant webpages. By contrast, content-based multimedia retrieval provides the possibility of analysing multimedia materials on the actual content. The availability of these technologies for multimedia search suggests further investigation of content-based hyperlinking for multimedia collections.

This thesis is dedicated to a novel topic of automatically creating hyperlinks within TV data collections for content-based browsing and navigation. Hyperlinks are created between video segments determined to be related based on their multimodal features.

First, we detail the methodologies to create potentially relevant segments across the TV collection in terms of automatically detected spoken information. We present which of these approaches are more efficient to segment video streams.

Next, we involve both low-level and high-level visual features to improve the hyperlinking quality. We detail the implementation of data fusion schemes to combine multimodal features.

Finally, a novel hyperlinking framework associated with query enrichment, spoken data analysis, and multimodal fusion is proposed. The experiments

show the effectiveness of this framework at satisfying user experience which is concluded in crowdsourcing study.

List of Publications

- **Shu Chen**, Kevin McGuinness, Robin Aly, Noel E. O'Connor, Tinne Tuytelaars. "*AXES @ TRECVid 2014: Instance Search*", In Proceedings of TREC Video Retrieval Evaluation, 2014.

- **Shu Chen**, Gareth J.F. Jones, Noel E. O'connor. "*DCU Linking Runs at MediaEval 2014: Search and Hyperlinking Task*", In Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, October 2014.

- Maria Eskevich, Robin Aly, David N. Racca, Roeland J.F. Ordelman, **Shu Chen**, Gareth J.F. Jones. "*The Search and Hyperlinking Task at MediaEval 2014*", In Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop, Barcelona, Spain, October 2014.

- **Shu Chen**, Maria Eskevich, Gareth J. F. Jones, Noel E. OConnor. "*An Investigation into Feature Effectiveness for Multimedia Hyperlinking*", In MultiMedia Modeling 2014, Dublin, Ireland. Pages 251-262, January, 2014.

- Maria Eskevich, Gareth J.F. Jones, Robin Aly, Roeland J.F. Ordelman, **Shu Chen**, Danish Nadeem, Camille Guinaudeau, Guillaume Gravier, Pascale Sebillot, Tom de Nies, Pedro Debevere, Rik Van de Walle, Petra Galuscakova, Pavel Pecina, Martha Larson. *Multimedia Information Seeking Through Search and Hyperlinking*, In Proceedings of the 3rd ACM conference on International Conference on Multimedia Retrieval (ICMR '13), Pages 287-294, 2013

- **Shu Chen**, Gareth J.F. Jones, Noel E. O'Connor. "*DCU Linking Runs at MediaEval 2013: Search and Hyperlinking Task*", In Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, October 2013

- Maria Eskevich, Robin Aly, Roeland J.F. Ordelman, **Shu Chen**, Gareth J.F. Jones. *The Search and Hyperlinking Task at MediaEval 2013*, In Proceedings of

the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, October 2013.

- Robin Aly, Roeland J.F. Ordelman, Maria Eskevich, Gareth J.F. Jones, **Shu Chen**. *Linking inside a Video Collection: What and How to Measure?*, In Proceedings of the 22nd International Conference on World Wide Web Companion (WWW '13 Companion), Republic and Canton of Geneva, Switzerland, Pages 457-460, 2013

- **Shu Chen**, Gareth J.F. Jones, Noel E. O'Conner. "*DCU at NTCIR-10 cross-lingual link discovery (CrossLink-2) task*", In Proceedings of NTCIR-10 Workshop Meeting, Tokyo, Japan, June 2013.

- Robin Aly, Kevin McGuinness, **Shu Chen**, Noel E. O'Connor, Ken Chatfield, Omkar Parkhi, Relja Arandjelovic, Andrew Zisserman, Basura Fernando, Tinne Tuytelaars, Dan Oneata, Matthijs Douze, Jerome Revaud, Jochen Schwenninger, Danila Potapov, Heng Wang, Zaid Harchaoui, Jakob Verbeek, Cordelia Schmid. "*AXES at TRECVID 2012: KIS, INS, and MED*", In Proceedings of TREC Video Retrieval Evaluation, 2012, USA.

- **Shu Chen**, Gareth J.F. Jones, Noel E. O'Connor. "*DCU Linking Runs at MediaEval 2012: Search and Hyperlinking Task*", In Proceedings of the MediaEval 2012 Multimedia Benchmark Workshop, Pisa, Italy, October 2012.

- Maria Eskevich, Gareth J.F. Jones, Robin Aly, Roeland J.F. Ordelman, **Shu Chen**, Danish Nadeem, Camille Guinaudeau, Guillaume Gravier, Pascale Sebillot, Tom de Nies, Pedro Debevere, Rik Van de Walle, Petra Galuscakova, Pavel Pecina, Martha Larson. "*Multimedia Information Seeking through Search and Hyperlinking*", In Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval (ICMR '13), Pages 287-294, 2013.

- Maria Eskevich, Gareth J.F. Jones, **Shu Chen**, Robin Aly, Roeland J.F. Ordelman, Martha Larson. "*Search and Hyperlinking Task at MediaEval 2012*", In Proceedings of the MediaEval 2012 Multimedia Benchmark Workshop, Pages 4-5, Pisa, Italy, October 2012.

- **Shu Chen**, Kevin McGuinness, Robin Aly, Noel E O'Connor, Franciska De Jong. "*The AXES-lite Video Search Engine*", In Image Analysis for Multimedia Interactive Services (WIAMIS), Pages 1-4, Dublin, Ireland, 2012

Abbreviations and Acronyms

ASR - Automatic Speech Recognition

BRIEF - Binary Robust Independent Elementary Features

DoG - Difference of Gaussians

ESA - Explicit Semantic Analysis

HSV - Hue-Saturation-Value colour model

IR - Information Retrieval

NE - Named Entity

NER - Named Entity Recognition

MDM - Maximum Deviation Method

ME13data - The TV video collection for the MediaEval 2013 Search and Hyperlinking Task

ME14data - The TV video collection for the MediaEval 2014 Search and Hyperlinking Task

ORB - Oriented FAST and Rotated BRIEF

RGB - RGB colour model

SIFT - Scale Invariant Feature Transform (SIFT)

SURF - Speeded-Up Robust Feature

SVM - Support Vector Machine

TRECVid - TREC Video Retrieval Evaluation

VSM - Vector Space Model

Contents

Chapter 1

Introduction

1.1 Overview

Information retrieval deals with the representation, storage, organisation, and access to items such as documents, web pages, online catalogues, structured and semi-structured records, and multimedia objects in order to satisfy a user's information needs [BYRN99]. Depending upon the different organisation and representation of information resources, there are two major strategies for seeking relevant items, search and navigation.

Search, as a well-known information retrieval (IR) strategy, is widely researched and developed for academic and commercial usage. Users input a query representing their information need. Depending on the search system and the information need, a query can be in the form of a text string or one or more images for multimedia IR. Typically, a query does not uniquely identify a single relevant object in the collection which is able to satisfy the search's information requirements. Instead, a number of objects may be identified as potentially relevant and returned to the user as the "search result". If necessary, the procedure of search can be iterated until a user's need is satisfied. In modern IR, search engines,

such as Google[1] and Yahoo[2], are well developed for search over content found on the World Wide Web, and achieve considerable success in both academic and commercial areas.

Nowadays, engineers are dedicated to designing powerful search engines to connect the resources on the World Web Wide, and researchers have proposed numerous algorithms to provide more accurate and efficient searching services. Even though such a large number of technologies have been developed to improve retrieval quality, in the context of increasing amounts of multimedia data, search alone can not satisfy user requirements for exploring multimedia resources due to reasons such as:

- The data in multimedia collections now is often very large and is increasing at very high rate. Taking an example of Youtube[3], according to a public statistic[4], there are a total of over 4 billion videos on Youtube, and on average, 300-hour videos are uploaded to the website per minutes. Search focusing on only user queries and top retrieved results will often not reflect the rich contents of potential interest to the user.

- Engineers expect that searching can present plentiful resources relevant to a query. However, users typically do not investigate more than a few retrieved items. They often read only a small number of top ranked results then either be navigated to interesting links or change queries to update retrieval results. The conflict between the searching mechanism and user performance suggests that focusing on top relevant documents in the retrieval list could satisfy users' requirements.

[1]http://www.google.ie
[2]http://ie.yahoo.com
[3]http://www.youtube.com
[4]http://expandedramblings.com/index.php/youtube-statistics/

- A well-designed search engine can locate relevant resources with a higher rank. When viewing only a few items, users are likely to find what they need. However, users vary in their ability to form efficient queries to make a clear expression of their searching requirement.

In conclusion, the effectiveness of search for information discovery in multimedia archives is inherently limited by the previous factors. Thus, navigation is used as a complementary mechanism to enrich user browsing experience on the World Web Wide. Navigation usually guides users from one resource to other relevant through hyperlinks. It satisfies spontaneous information needs of users when they inspect the content of resources [FHHD92] by: 1) manually or automatically preprocessing a large data collection to provide a meaningful roadmap to enrich user's browsing experience; 2) guiding users from one resource to other top relevant resources; 3) allowing users to visit potentially relevant resources without any input.

In computer science, a hyperlink is a common implementation of navigation by creating a reference from which a user could navigate other local or online resources. A common sample of a hyperlink is linking a word or multiple words within a webpage to other relevant documents according to its semantic information. For example, in Wikipedia[5], a webpage titled with "Web Search Engine" defines itself as "A web search engine is a software system that is designed to search for information on the World Wide Web", where the words "World Wide Web" are linked to another page titled with "World Wide Web". A reader can be navigated to this page and fully understand the topic of "World Wide Web".

Navigation, utilising hyperlinks in the World Web Wide, enables a user to access a linked target item which is expected to be relevant to or related to the user's interests or worthy of recommendation to the current user according to his/her browsing context. It can reduce the negative influence of poor user

[5]http://www.wikipedia.org

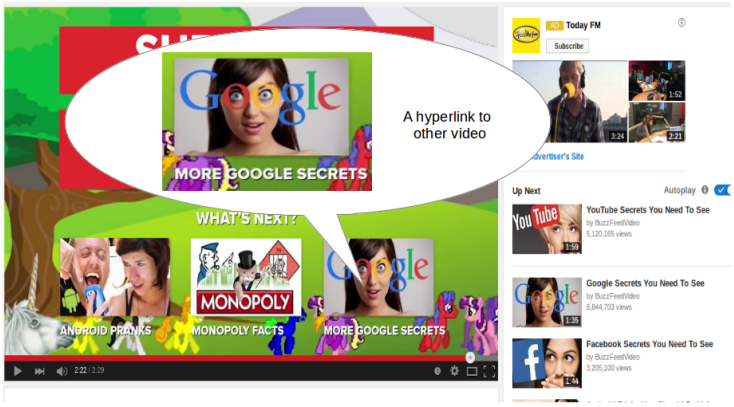
searching skills since content creators or curators can maintain the data and its corresponding linked targets manually. Users can begin seeking data with a query-based search to identify a relevant starting point to explore potentially interesting resources. Nowadays, user navigation is no longer limited to text-based webpages. Multimedia resources, especially videos, are widely applied for modern information interactivity. Compared with classic text-based documents, video archives typically contain a combination of features described by the audio track and visual content. Creating hyperlinks within such a rich content associated with video archives requires further investigation.

1.2 Research Motivation

Nowadays, multimedia hyperlinks are commonly implemented on video-watching websites. Taking the example of Youtube[6], we outline two different types of multimedia hyperlinks, illustrated in Figure 1.1. Figure 1.1 (a) illustrates video recommendation via automatically created video-to-video hyperlinks based on usage recommendation. Both of them preview the videos sharing the same topic "Star Wars". Figure 1.1 (b) presents an anchor-to-video hyperlink which navigates users to other potentially interesting videos on Youtube. The hyperlink locates in the video to present Google research in information retrieval and guide users to another video introducing more about the Google company. The manual construction of links is an obvious difference from the previous example. These two examples represent state-of-the-art use of hyperlinks in existing multimedia systems. Usually, the relevance analysis between anchors for the hyperlink is determined based on manually generated tags, video titles and the corresponding

[6]http://www.youtube.com

(a)



(b)

Figure 1.1: Example of multimedia-based hyperlinks

brief introduction. However, there are a number of disadvantages brought by these kinds of hyperlinks.

- A video-to-video hyperlink can construct a robust content-based link between short videos since a short video is typically focused on a single topic. Given this, hyperlink creation can be regarded as a topic-to-topic matching process. A long video could contain multiple segments whose topics vary. Users navigated by the video-to-video hyperlink are expected to browse the relevant topic shared by the two archives. However, various or even

irrelevant topics potentially existing in a target video could cause difficulty for users to accurately find the relevant part.

- Links between manually created anchors and targets can produce an accurate correlation between the linked targets and decrease potential content-based mismatch between linked resources. However, creating hyperlinks manually across a large multimedia collection requires exorbitant expenditure considering the huge size of online data collections and their rapid growing speed.

Our research applies the anchor-to-target multimedia hyperlinking framework. The objective is to investigate how multimodal features perform in this hyperlinking framework. The primary difference of an anchor-to-target hyperlink across video archives compared with other cases, for example a classic IR system or a recommendation system, can be understood as "give me more information about this anchor" instead of "give me more based on this anchor or entity" [OEA+15]. It simulates a scenario where a user browses an audiovisual archive and accidentally finds something interesting. The user would like to browse other video archives relevant to the current interesting point. An anchor-to-target hyperlinking system can be applied by receiving an anchor as a query for the content-based multimedia IR system. The system utilises information retrieval techniques to identify relevant fragments in the video archives and finally presents these results to users.

A primary motivation of this thesis is integrating classic multimedia IR techniques to construct a multimedia hyperlinking system. We investigate the optimal solutions for various multimodal feature analysis, examine different hyperlinking methodologies in terms of human perspective, and propose an efficient anchor-to-target hyperlinking framework.

1.3 Research Objectives

According to the proposed hyperlinking systems in the previous discussions, we make the research hypothesis: 1) hyperlinking systems in video collections could improve users' browsing experience; 2) an implementation of this system should involve multimodal features to reflect users' cognition when watching videos. Thus, the research objectives of this thesis are:

- First, we review a review of state-of-the-art investigations in anchor-to-target multimedia hyperlinking. By comparing their achievements and deficiencies, we propose the research questions in building multimedia hyperlinking systems.
- Second, we start our research from text-based IR techniques. A comparison between various methodologies for hyperlinking segmentation is presented based on the textual content. The results are used as baselines for further investigation.
- Third, we extend the previous study to utilise visual features to investigate the usefulness of visual information in video hyperlinking. Experimental results will show how individual visual features influence hyperlinking performance.
- Finally, we investigate different methodologies to integrate multimodal features using data fusion.

1.4 Thesis Structure

This remainder of this thesis is organised as follows:

Chapter 2 reviews existing research on multimedia hyperlinking, content-based multimedia information retrieval, and other relevant topics. First, we

present a review of content-based information retrieval. Next, we outline state-of-the-art approaches to integrate IR results from different retrieval systems that can be applied to the remainder of this thesis. Finally, we review a detailed history of the hyperlinking task from text-based to multimedia-based.

Chapter 3 introduces the hyperlinking framework adopted in this thesis, research questions to be addressed, and experimental hypothesis used in Chapter 4, 5 and 6.

Chapter 4 focuses on text-based hyperlink generation. Hyperlinks are created according to the corresponding spoken information extracted by automatic speech recognition algorithms. Two classic text information retrieval models, TD-IDF and Okapi BM25, are used to index and search related segments. The experiment reveals the superiority of Okapi BM25 when its parameters are suitably adjusted for the hyperlinking task.

Chapter 5 introduces various multimodal features, including the colour histogram, the Bag-of-Visual-Words model based on a low-level descriptor, semantic concepts and manually generated metadata information. We compare hyperlinking results retrieved by each modality and utilise a fusion scheme to combine multimodal results.

Chapter 6 investigates various methodologies to generate a text-based query for multimedia hyperlinking. We demonstrate that this strategy can significantly improve hyperlinking performance. Moreover, we propose a final hyperlinking framework using query content analysis, multimodal feature retrieval, and our fusion scheme.

Chapter 7 concludes the thesis and proposes areas for future investigation.

Chapter 2

Literature Review

2.1 Chapter Overview

This chapter reviews relevant literature on multimedia hyperlinking research. Section 2.2 presents and compares existing research on content-based multimedia processing including spoken data analysis, visual descriptor extraction and the creation of semantic concepts using multiple low-level multimedia features. Section 2.3 introduces multimodal feature fusion techniques, including an overview of two main fusion schemes: early fusion and late fusion, in which we discuss the effectiveness of both methods for multimedia analysis. Section 2.4 gives an overview of hyperlinking system development. Section 2.5 concludes this chapter.

2.2 Multimedia Features for Content-based IR

Multimedia hyperlinks connect video segments across multimedia collections based on the content of the segments. Given a video segment, a hyperlinking system retrieves segments and shows the most relevant ones to users. In essence, the retrieval process during multimedia hyperlinking is an automatic multimedia retrieval. Content-based multimedia analysis can be applied to hyperlink con-

struction. To this end, multimodal features, including video descriptions, audio tracks, visual content, can be extracted from multimedia resources and used in link construction. These multimodal features, represented in various formats, can be broadly categorised into two types: low-level and high-level features. An overview of these features is introduced in the following sections.

2.2.1 Low-level Multimodal Features

Low-level features are extracted directly from the digital representation of the multimedia content. [XZT+06] described low-level features as those image/video features that can be easily extracted to represent colour, texture, shape, or audio characteristics of multimedia content, thus most of them only reflect basic textual or visual features rather than human perception. In this thesis, we consider two categories of low-level features: text-based and visual-based.

A text-based low-level feature converts the audio track into text words to represent a video story. Human annotation can provide a description of an audio track by creating time-based subtitles. We can expect very high quality of subtitles when produced by professional workers albeit with a relatively high associated cost. An alternative approach is to extract spoken information automatically. Automatic Speech Recognition (ASR), a computer driven mechanism to translate spoken language into readable words [Stu94], can often be applied at relatively low cost.

Another type of low-level feature is constructed based on the visual information. These visual features are detected from the keyframes extracted from a video stream. [Goo00] provided a further consideration that most algorithms to detect low-level visual features focus on three aspects of image characteristics: colour, texture, and shape. In this thesis, we use the colour feature as one primary low-level visual descriptor.

Colour Features

Colour features are widely used in content-based multimedia retrieval. [VR04] concluded that colour features are extensively used in image database retrieval due to their robustness to noise, resolution, orientation and resizing. [YH07] confirmed that colour features provide strong cues that capture human perception in a low dimensional space, and that they can be generated with less computational effort than other advanced features.

A colour feature characterises images in terms of the colour space that describes a specific organisation of colour information. The following introduces two classic colour spaces: the RGB colour model (RGB), and the Hue-Saturation-Value colour model (HSV).

RGB utilises a three-dimensional vector to represent human perception of natural colour. Researchers can derive other kinds of colour representations by using either linear or non-linear transformations from the RGB model [CJSW01]. Due to the simple implementation and efficiency of visual representation, the RGB model is often used as the fundamental colour space in content-based multimedia retrieval [SWS+00, SQP02b, WCL07, DJLW08, SHP12].

[PKPM09] outlined a disadvantage of RGB as the fact that the distance computed between two colours in RGB space may not reflect their perceptual similarity. In [Smi78], the authors introduced the perceptual properties of "hue", "saturation,", and "value" to approximate human concepts on the colour space, as the HSV model. HSV offers an intuitive and perceptual representation of colour information by mapping the values into a cylinder[Lin12]. Existing research [ORMA01, Ma09, SQP02b, KB13, LS13] has shown the effectiveness of HSV. [ORMA01] compared HSV and RGB, and concluded that HSV achieved the best retrieval quality when applied to multiple image databases provided by Corel-

GALLERY[1], QBIC Developers Kit CD-ROM[2], and Swedish University Network FTP Server Images[3]. [SQP02b] demonstrated the outstanding performance of HSV over 14,500 images collected from the Internet.

A colour space utilises a histogram to describe colour distribution with an N-dimensional vector. The similarity score (or distance) between two colour histograms can be represented in a quadratic form [HSE+95]. The primary contribution of similarity metrics is the use of a low dimensional vector to measure the colour distribution. Existing similarity metrics for colour histograms include the cosine distance [SQP02a], Euclidean distance, Chi-square kernel [PW10], Bhattacharyya distance [CRM03], Convolution Kernel, Correlation, etc. In the thesis, we use the Correlation which is supported by OpenCV[4]. Its efficiency has been demonstrated in [CEJO14, CJO12, CJO13] to measure the colour distribution. Further details are provided in Chapter 5.

Visual Descriptors

Low-level features are not restricted to colour. Research in [FFP05] outlined that using an intermediate representation can achieve better content-based image retrieval, and that these intermediate representations can be a mixture of textures or codewords. It raised a research question of how to interpret an image by using intermediate visual features. The Bag-of-Visual-Words (BoVW) model explains the visual content by transferring intermediate visual features to a vector of "words" which maps each feature to an occurrence of an entry in a visual vocabulary. An initial motivation of BoVW was to construct a vocabulary of prototype tiny surface patches with associated local geometric and photometric

[1]http://www.corel.com
[2]http://wwwqbic.almaden.ibm.com
[3]ftp://ftp.sunet.se/pub/pictures
[4]http://opencv.org/

properties [LM01]. [FFP05] further outlined the steps to implement BoVW as: feature detection, feature description and vocabulary generation.

Feature detection interprets the content of an image as a set of intermediate features. Researchers are dedicated to investigating effective approaches to match images since [Mor81] in which a corner detector was used to match images. [HS88] improved the mechanism of corner detection and proposed Harris corner detector to be widely used in image matching tasks. However, [Low04] pointed out that Harris corner detector is sensitive to scaled images. To address this issue, the author in [Low04] proposed an approach to create scale- and rotate-invariant visual descriptors, referred as to Scale Invariant Feature Transform (SIFT). This overcomes the disadvantage of Harris corner detector by generating scale-invariant interest points from an image.

Each image (or an interest area) is represented by a set of N keypoint descriptors which contain a total of 128 bin values (there are 16 blocks of size 4×4 and each block contains 8 bin orientation histogram). Therefore, SIFT utilises a $N \times 128$ matrix to describe interest points in each image (or an interest area). The advantage of SIFT is its distinctiveness. This is achieved by assembling a high-dimensional vector representing the image gradients within a local region of the image, which enables the correct match for a keypoint to be selected from a large database of other keypoints [Low04]. Since its introduction, SIFT has found widespread application in computer vision and content-based information retrieval [SZ03, SZ06, KS04, RRKB11].

Researchers investigated a number of novel image descriptor schemes since the successful introduction of SIFT, and developed a number of descriptors. [BTVG06] proposed the scheme "Speeded-Up Robust Feature" (SURF) which accelerates the process of detecting potentially interest points and creating visual descriptors with the advantage of rotate- and scale-invariance. Furthermore, according to [BTVG06], SURF outperformed SIFT both in speed and accuracy. The authors of

[RD06] introduced a high-speed corner detection algorithm named "Features from Accelerated Segment Test" (FAST), which achieves faster corner detection than other algorithms, including Harris Corner Detection or Difference of Gaussians. [CLSF10] improved the memory consumption of SIFT and presented "Binary Robust Independent Elementary Features" (BRIEF). This scheme uses smoothened image patches and computes binary strings from them, instead of floating-based descriptors that are applied in SURF and SIFT. The experimental investigation using BRIEF in [CLSF10] showed that it outperformed SURF. [RRKB11] integrated FAST and BRIEF to create a novel scheme "Oriented FAST and Rotated BRIEF" (ORB), which inherited the efficient corner detection of FAST and the accuracy of BRIEF.

Image descriptor schemes interpret each image as an $N \times M$ matrix in which N is the pre-defined number of interest points, and M is the dimension of descriptors. A visual vocabulary is then defined according to the distribution of the descriptor matrix. A state-of-the-art approach to create this vocabulary is to cluster the descriptor matrix extracted from the image collection and define the centres of the learned clusters as the visual vocabulary. K-means is widely applied to implement descriptor clustering [LM01, SZ03, SZ06], although, [SZ03] suggested some alternative approaches including K-medoids and histogram binning.

An open issue when creating a vocabulary for a multimedia collection is how to determine the size of the visual vocabulary. After [SZ03], researchers investigated vocabulary size to improve content-based multimedia retrieval. In [SZ03], the size of K-means cluster centres was set to be 6,000 and 10,000 respectively. The authors of [PCI+07] regarded the visual vocabulary as a primary computational bottleneck. To address this issue, they examined different scalable approaches to determine visual vocabularies. The vocabulary size was set to be 50K, 100K, 250K, 500K, 750K, 1M and 1.25M. The experimental investigation on the Flickr[5] collec-

[5]http://www.flickr.com/

tion revealed that a large visual vocabulary benefits content-based information retrieval and that a vocabulary of 1M was optimal.

Low-level visual features have been applied into various multimedia retrieval tasks. Take some examples: [SZ03] implemented a well-know video search engine, Video Google, involving SIFT descriptors; [BBL+08] used RGB colour space to summarise BBC TV collections in TRECVid 2008; [SGF+11] applied SURF descriptors in Know-Item Search task in TRECVid 2011; [AMC+12] used SIFT descriptor as the primary feature in multimedia event detection task, know-item search task, and instance search task. In our hyperlinking research, matching visual content is an important stage to determine the relevance between two linked video clips. We involve the introduced low-level features, colour spaces and visual descriptors, as two primary multimodal features to be investigated. One of the motivations is the rich research experience of using low-level features in multimedia retrieval.

The other motivation for using low-level visual features is comparing their hyperlinking results with that of high-level visual features. Researchers already learn that low-level features sometimes fail to satisfy user requirements [CH05], due to the lack of coincidence between the information contained in visual data and the interpretation given by users for the same data [ZLS+06] [SWS+00]. Therefore, high-level features are developed to describe the semantic information in images. We expect to involve both low-level and high-level features to investigate their difference in hyperlinking retrieval. In next section, we review the topic of high-level features.

2.2.2 High-level Multimodal Features

High-level representation techniques are based on the idea of recognising models of objects presented in an image and identifying image regions representing

human cognition of visual content [Tam08]. The authors of [LHR99] suggested that high-level features, also named conceptual features, must be based on low-level features. Before 2000, researchers focused on how to extract high-level features by using geometric elements, such as point set [Fau93], shape description [LHR99], or contour segmentation [NB80]. The following part will briefly review some relevant works.

In digital images, a point set can be regarded as pixels [Fau93], feature vectors [DHS73], objects [Oga86] or spatial relationships. There are several classic algorithms to detect point sets for high-level feature generation, including border tracking, Hough Transform (HT), etc. Among them, the HT, according to [DH72], was widely to find imperfect instances of objects within a certain class of shapes by a voting procedure. The HT is widely applied in detecting arbitrary geometric shapes, like circles or ellipses.

Shape description uses geometric elements, including the total number of pixels, length, perimeter, compactness, or topological description, to represent the outer shape of objects in digital images [LHR99].

Contour segmentation is designed to outline the outer shape of objects accurately in digital images. The authors of [ZVC89] proposed the idea of edge linking and segmentation, which emphasised identifying local edge pixels, linking them to contours and segmenting contours.

After 2000, the definition of high-level concepts bridges visual content recognition with user requirement in IR systems. Research suggested high-level concepts should present not only the objects contained in multimedia resources but also those which are potentially attractive to users. [Pet00] defined the need for content-based information as responding to querying the content of multimedia resources. The authors advanced a structure of multimedia retrieval, which firstly inferences high-level concepts learned from multimodal features as queries, and then searches relevant documents in multimedia collections.

[LLYK06] presented the development of a movie skimming system to create video abstraction. The structure used both low-level and high-level features to improve video summarisation performance. [TV07] analysed how video abstraction techniques facilitated the requirement for browsing digital video infrastructures. [Sme07] purposed several video retrieval strategies as follows: 1) using metadata and browsing keyframes; 2) using text for video searching; 3) keyframe matching; and 4) semantic features for video retrieval. Among them, semantic features and metadata can both be categorised as high-level features.

[XZT+06] integrated different types of videos, which are relevant to online meeting, movies, broadcast news, and sports, into a general framework. The authors pointed out a critical issue in content-based video retrieval is decreasing the gap between the description of objects in human observation and in computational representation. In computer science, this difference between two descriptions of an object by different informative resources is "semantic gap" [SWS+00]. The semantic gap is not the patent of high-level visual concepts, while integrating information from different resources could cause it. In this thesis, the semantic gap, however, refers in particular to the gap between high-level concepts and low-level descriptors when describing an object in multimedia resources.

Over the last few years, researchers have tried various strategies to overcome the semantic gap. Manual indexing of multimedia resources is an effective methodology to facilitate multimedia IR [SOK09]. One example of manual indexing is manually creating an abstract of a video. However, this approach often lacks the detailed representation at the shot level and needs a significant cost [SOK09]. Another approach to decrease the semantic gap is increasing the size of the concept collection. [HYL07] examined the optimal concept size and confirmed that a few thousand semantic concepts could be sufficient to support high accuracy video retrieval in TRECVid collections. Besides, researchers have investigated various approaches to improve the quality of high-level concept annotation. [SF10]

used Support Vector Machine (SVM) to annotate and tag interest instances in images, and suggested that the structured SVM [JFY09] outperforms normal SVM. [JZCL08] investigated how to apply SVM for annotating concepts across various data collections, which was defined as cross-domain learning. They developed cross-domain SVM (CDSVM) and showed the superiority of this algorithm using TRECVid data collections. Recently, deep learning is a popular area of Machine Learning research and multiple research [SSZ12, CSVZ14, Le13] applied it for semantic concept recognition.

Researchers have sought to identify sets of high-level concept collections for content-based multimedia IR. We briefly review two cases. The first one is Object Bank [LSFFX10] provided by Visual Lab, Stanford University. It contains a total of 177 high-level concepts created by pre-trained generic object detectors. Each keyframe is described as a feature vector which is calculated using a three-level spatial pyramid (1×1, 2×2, 4×4) [LSP06]. The size of the feature vector for each image is 44,604. The strategy of Object Bank has inspired multiple research [SC12, ASD12, LSLFF14] in high-level concept construction. The second one is provided by the Vision Group at University of Oxford according to [CLVZ11], specially created for the MediaEval Search and Hyperlinking task [EJC$^+$13, EAL12]. The collection contains a set of concept detector scores for 1,000 concepts. The detectors were trained using on-the-fly concept detection approach proposed in [CZ13], which obtains relevant images from Google as the training collection, and learns the difference using SVM classifier. This collection will be used for the experimental investigation in the remainder of this thesis.

2.3 Data Fusion

In [SW05]'s perspective, the video content can be expressed using varying information including visual, auditory, or textual features. However, multimodal

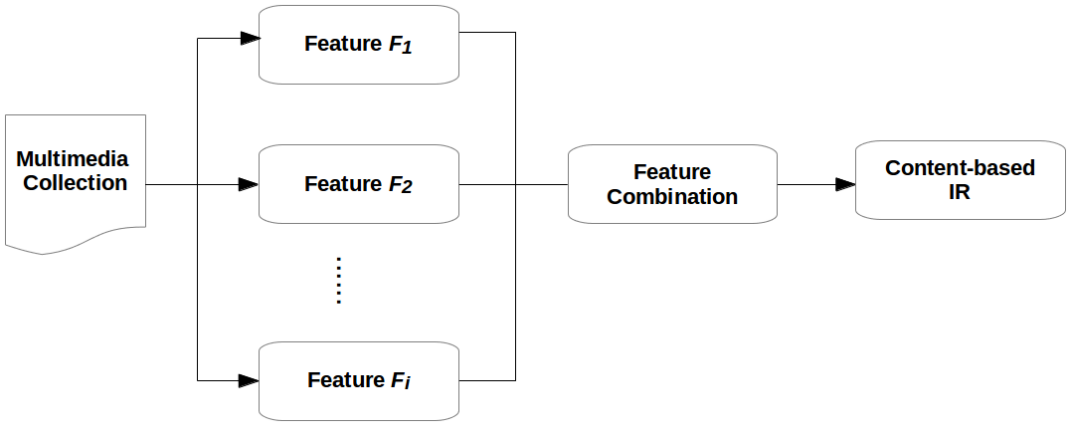


Figure 2.1: An early fusion scheme

information retrieval in early years lacked theories to answer what are the best multimodal feature to achieve the best retrieval results [WCCS04]. With the increase of experimental investigations, researchers realised that multimodal features are inherently noisy, and the retrieval results from separate IR systems applying an individual feature are potentially unreliable [Wil09]. Thus, in content-based multimedia information retrieval, various research [AHESK10, WZZL10] tried to describe the document in multiple perspectives by using data fusion to integrate different multimodal features. Data fusion aims to improve retrieval quality by combining retrieval results from multiple IR systems to produce a new and hopefully better ranking [FV07].

2.3.1 Fusion Schemes

There are two major data fusion schemes used in multimodal information retrieval, early fusion and late fusion, which are distinguished by the strategy of combining feature analysis results of individual information retrieval procedures. They are both defined concisely in [SWS05]. Early fusion integrates multimodal features before learning concepts. The early fusion process is shown in Figure 2.1.

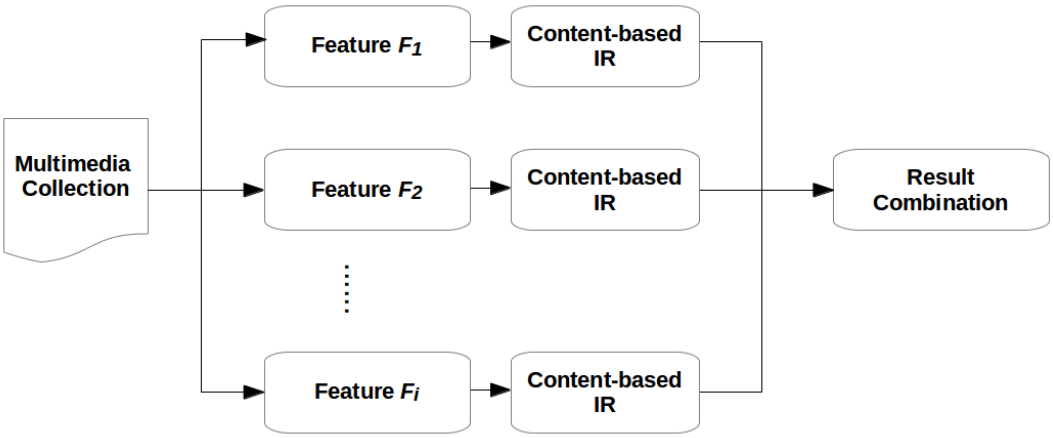


Figure 2.2: A late fusion scheme

Multimodal features are extracted from the multimedia collection. These features are integrated to generate a combined feature to represent a new description of its content. Content-based IR utilises the newly generated features to collect the retrieval results. Late fusion firstly applies multimodal features to retrieval the ranked lists from different system separately, and then these scores in the ranked lists are integrated to produce a final retrieved list. The late fusion process is shown in Figure 2.2.

[SWS05] concluded that the challenge of early fusion is how to integrate multimodal features into to a common representation, while the late fusion scheme fuses the retrieval scores rather than a combined feature representation [SWS05]. However, an issue to be addressed in the late fusion scheme is the combination strategy used. In the following section, we introduce a widely applied late fusion scheme - linear data fusion.

2.3.2 Linear Data Fusion

A linear data fusion can be represented as:

$$\text{CombSUM} = \sum_{i=1}^{K} w_i \cdot R_i, \tag{2.1}$$

where R_i denotes the ranked list from the ith IR system, K is the total number of fused IR systems, and w_i is the fusion weight for the ith retrieval result. Equation 2.1 shows the standard CombSUM model introduced in [FS94]. CombSUM calculates the fused score by adding the weighted score of a document in each of the result lists. Another data fusion scheme defined in [FS94] is CombMNZ. It extends CombSUM by introducing a variable $n(d)$ which weights the element based on how many result sets a document d appears in [Wil09], shown as Equation 2.2. Existing research has utilised several varieties on the CombSUM scheme to implement linear combination. [SSH14a] proposed linear fusion as Equation 2.3, which implemented binary feature fusion between different IR systems. A widely applied linear fusion to integrate binary features is defined as shown Equation 2.4 according to [MLD+14, VC99, CEJO14], which can be regarded as a simple version of Equation 2.1.

$$\text{CombMNZ} = \text{CombSUM} \cdot n(d), \tag{2.2}$$

$$\text{Score}_{\text{fuse}} = {R_1}^{w} + {R_2}^{(1-w)}, \tag{2.3}$$

$$\text{Score}_{\text{fuse}} = w \cdot R_1 + (1-w) \cdot R_2, \tag{2.4}$$

The fused result $\text{Score}_{\text{fuse}}$ is calculated by combining the normalised score of a document d in the corresponding results retrieved by the IR system R_i. Therefore, score normalisation is a key issue that determines the quality of linear fusion. According to [MLD+14], score normalisation algorithms include MinMax and rank-based normalisation. MinMax normalisation linearly transforms the retrieved score according to:

$$\text{Score}_{\text{linear\_normal}} = \frac{\text{S}_{\text{retrieval}} - \text{S}_{\text{min}}}{\text{S}_{\text{max}} - \text{S}_{\text{min}}}, \tag{2.5}$$

where $\text{S}_{\text{retrieval}}$ is the retrieved score in a retrieval collection, and S_{min} is the minimum score in this retrieved collection, and S_{max} is the maximum score. Rank-based normalisation can be defined according to Equations 2.6 and 2.7:

$$\text{Score}_{\text{rank\_normal}} = N - R(d), \tag{2.6}$$

$$\text{Score}_{\text{rank\_normal}} = \frac{1}{R(d)}, \tag{2.7}$$

where N is the total number of retrieved documents, $R(d)$ returns the rank position of a document d in the retrieval system R_i. Rank-based normalisation usually produces the same similarity score for the document at the same rank in different ranked lists. In our research, we expect that the score difference between multimodal hyperlinking systems using different features can reflect the importance of the corresponding features. Therefore, the experimental results in the remainder of this thesis focus on MinMax normalisation. In Chapter 5, we introduce an estimation of fused scores based on the theory of rank-based normalisation.

Another open issue for the linear fusion scheme, according to [AHESK10], is the determination of the fusion weights. A widely applied approach is to use equal weights for each merged IR system, meaning that the variable w_i in Equation 2.1 is set to 1. This approach assumes that each IR system contributes equally to producing good results. Using equal fusion weights requires no more investigation, and this methodology is regarded as a low-cost implementation. However, research [MLD+14] showed that estimating weights can improve the data fusion results in multimedia hyperlinking. Thus, in this thesis, using equal weights is usually used as a baseline for our further investigation.

Approaches to estimating the weights of combining different IR systems can be categorised as supervised and unsupervised mechanisms. Supervised mechanisms optimise the fusion weights based on training data and then apply these optimised results on test or operational data. The criterion for determining the most efficient fusion weights is usually an evaluation metric for the performance of an IR system, for example, in [MBL+10], Mean Average Precision (MAP) was applied to optimise fusing weights. A grid search strategy was applied to optimise fusion weights on the training collections according to [SSH14b, MBL+10, INN03]. This enumerates a set of potential weight values for each IR system and examines the retrieval quality using the evaluation metric. Often, the grid search strategy is only used to determine the fusion weights between two IR systems due to its computational complexity. In [MLD+14], an alternative approach was proposed by using Fisher Linear Discriminant Analysis. Instead of achieving a numerical optimisation, this algorithm uses the optimal linear combination of multiple IR system. The coefficients of the optimal linear combination were regarded as the corresponding fusion weights. The authors in [MLD+14] demonstrated its effectiveness using the ImageCLEF collection [TK09]. The unsupervised mechanism to estimate fusion weights can be attractive based on the fact that no training set is required. [Wil09] presented the Maximum Deviation Method (MDM) approach and achieved better content-based multimedia IR performance on TRECVID 2003 and 2004 test collections. We will introduce a detailed description of this approach in Chapter 5.

2.4 The Emergence of Multimedia Hyperlinking

This section reviews the research track in multimedia hyperlinking. We focus on the conferences/workshops in recent years in which the topic of "multimedia hyperlinking" was investigated. The following section will outline important

literatures during the emergence of hyperlinking investigation, and discuss how these inspire and contribute the experimental investigation in the remainder of this thesis.

2.4.1 Early Stages

The origin of the concept of a "hyperlink" can be tracked back to 1945: [Bus45] assumed a microfilm-based machine which could create a trail to link any two pages containing related information. Project Xanadu[6], found in the 1960s, first proposed the word hyperlink in their hypertext project inspired by the assumption in [Bus45]. The target of this project was improving WWW. Currently, web pages on the WWW utilise hyperlinks to facilitate users to browse a large number of online documents.

A common format of hyperlink is the underscored blue text in webpages. Users can be navigated to other webpage by a single-click on it. In Chapter 1, we illustrated different type of hyperlinks between video clips in Youtube[7]. The various type of hyperlinks motivates researchers to investigate how to create hyperlinks to different multimedia resources. A popular data collection for hyperlinking investigation is Wikipedia[8], supported by the non-profit Wikipedia Foundation, provides the largest online encyclopedia with free content [wik09]. It is essential to utilise the hyperlinks to search and browse such a rich online document set. Although manually created hyperlinks contribute Wikipedia's daily operation, researchers have already been aware of the potential value of automatic hyperlinking strategies for Wikipedia.

Mihalcea and Csomai presented a link system Wikify! [MC07] based on Wikipedia resources. The purpose of Wikify! is to link entities (textual words) using Wikipedia as the target knowledge based, named as automatic text wikifica-

[6]http://www.xanadu.com/
[7]www.youtube.com
[8]http://en.wikipedia.org

tion. Automatic text wikification requires solutions for two main tasks: automatic document keyword extraction to detect valuable entities and word sense disambiguation to determine which Wikipedia pages should be linked from an entity [MC07]. The system combined the two tasks to provide a rich text annotation service. The authors declared that Wikify! could improve user experience on the Internet by automatically enriching online documents, benefiting students by providing a convenient gateway to other encyclopedic information, and contributing new solutions to rich text annotation [MC07].

Wikification using the techniques described in [MC07] was not perfect at hyperlink detection and disambiguation. On one hand, topic indexing needs to parse all Wikipedia documents, which is computationally expensive. On the other, Wikify! only considered the probability of an entity to link to another document, without involving the context information before and after this entity. This strategy always constructed links whenever a possible anchor exists in other documents. An alternative approach to link creation based on Wikipedia pages was introduced in [MW08]. This approach uses machine learning algorithms to analyse the context information of an entity for word disambiguation. For a word with multiple semantic definitions, the system [MW08] could make a better prediction of its actual meaning in the current document, and create hyperlinks to the correct resources.

[BHdR11] presented work on linking multimedia resources for unskilled users. This linking system was constructed especially for news, multimedia and cultural heritage archives. The linking task was defined as linking items with a rich textual representation in a news archive to items with sparse annotations in a multimedia archive, where items should be linked if they describe the same or a related event [BHdR11]. [KG10] treated linking as an alignment task, which meant identifying items in a collection that discusses the same person, entity or concept [BHdR11].

[ACD+98] defined another area of linking rich textual documents as topic tracking, where items were connected when they discuss the same and related events.

Researchers' interest in hyperlinking investigation is not limited to processing Wikipedia or other online textual documents. They are also dedicated to building hyperlinking systems in video collections. [Dak99] concluded a set of video hyperlinking systems in early stages, including The Aspen MovieMap [Wal80], Video Finger [Wat89], Elastic Charles [BD90], HyperPlant [TYT+92], etc. This thesis also proposed an automatic tool for creating hyperlinking video. The author applied colour, motion, and texture features to detect video stories and objects in videos, and the hyperlinks were constructed with the detected evidences. [DACBJ99] proposed a novel video hyperlinking system using multiple features, including colour, texture, motion, and the position of objects. Using these features, the authors implemented a video hyperlinking interface supporting user interaction. Through this interface, users could indicate a specific object in the video and expect a retrieval of other relevant videos. [CdCC+05] presented a video hyperlinking system to provide interactive hyperlinks across TV programs. TV programs are linked at shot level and the system can deliver multiple contents including video metadata, video streams, etc. [TNW08] investigated a video hyperlinking system in which the hyperlink is determined at shot level as well. The aim of [TNW08] was not to demonstrate the effectiveness of different video features, but how to use random walk algorithms to improve linking between keyframes. Thus, we focused on the proposed structure of hyperlinking system: a hyperlink exists between two keyframes of the corresponding video shots.

In conclusion, research in hyperlinking has a long history since the origin of the "hyperlink". Researchers investigated various approaches to creating hyperlinks across different data resources, including both textual pages and video collections. In the next part, we review some hypelinking tasks which inspire the design of hyperlinking research.

2.4.2 NTCIR: Text-based Hyperlinking on Wikipedia

NTCIR-9 [SJ11] and NTCIR-10 [JS13] proposed the hyperlink creation task based on a Wikipedia documents as Cross-Lingual Link Discovery (CLLD) task. The CLLD task[9] aimed to detect potentially important semantic links between documents in different languages. The reason of discussing the CLLD task is that its evaluation methodology inspires our approach to build the ground-truth collection for hyperlinking evaluation.

The NTCIR CLLD task proposed a set of evaluation benchmarks to compare participants' submissions. [TIG+11] described that the evaluation metrics of CLLD task used Precision@N, R-Prec, and Mean Average Precision, which were fundamental metrics for IR evaluation. Furthermore, [TIG+11] presented the evaluation methods can be categorised as file-to-file and anchor-to-file. To build the ground truth for the CLLD task, [TIG+11] proposed two methodologies using Wikipedia Ground-Truth Run and Human Assessors. The former utilised existing links in the testing Wikipedia collections, which were deliberately removed before being published to the task participants. The submitted hyperlinks are judged as relevant only when the two linked entities (n-gram words) exist in Wikipedia pages. Human annotation was used for the second approach. Task organisers hired oversea students with bi-lingual professional skills. These students reviewed the linked results and indicated their relevance [TIG+11]. Those results manually judged were collected as the ground-truth collections and delivered for task evaluation and further investigation.

The video collection used in our research contains no hyperlinks between video clips. It means that before experimental investigation, we should build the ground-truth collection from the video resources. The strategy in CLLD, hiring students to manually annotate the ground-truth, inspires us: we can use

[9] http://ntcir.nii.ac.jp/CrossLink/

human intelligence to determine the relevance of linked video clips. In Chapter 3, we introduce how we apply crowdsourcing platform to manually build the ground-truth collection.

2.4.3 MediaEval Workshop: Searching and Hyperlinking Task

In 2012, the MediaEval workshop organised a Brave New Task termed Search and Hyperlinking Task, MediaEval 2012. Later, the Search and Hyperlinking Task became the primary task in MediaEval 2013 and 2014. The research proposed in this thesis is fully based on the hyperlinking mechanism designed in MediaEval workshop. Thus, this section reviews the corresponding publications and introduces how these works inspire experimental investigations in this thesis.

MediaEval 2012

MediaEval 2012 proposed a Search and Hyperlinking Task as a Brave New Task. This task was driven by the following use-case scenario: a user is searching for a known segment in a video collection, and on occasion the user may find that this segment is not sufficient to address their information need or they may wish to watch other related video segments [EAL12]. The Search and Hyperlinking Task required task participants to search for a known relevant segment and create hyperlinks to related video segments.

A total of four groups participated in the Hyperlinking subtask in MediaEval 2012. The authors in [DNDVD$^+$12] utilised a traditional Vector Space Model (VSM) and TF-IDF algorithm to index and search spoken transcripts. The results of named entity extraction were used to create the VSM and the similarity was calculated based on cosine similarity. In [GGS12], research focused on analysing various spoken transcripts using BM25 and VSM. Name entities for the document

vectors were extracted using TreeTagger[10]. Furthermore, the metadata of the videos was used to re-ranking the hyperlinking results. In [NAO12], the participants described the video content by high-level concepts. A total of 508 concepts were included based on their previous TRECVid investigation [CLVZ11]. We proposed our research in [CJO12] in which both textual and visual features were used. We use the TF-IDF algorithm to index and search the spoken transcripts and the bag-of-visual-words model to analyse the low-level visual features. The visual descriptors were extracted using the SIFT algorithm. The final hyperlinking results was determined by the combination of hyperlinking results using textual and visual features.

The experimental investigation revealed that the spoken information extracted by the ASR algorithms achieved better results, while only visual features, either high-level [NAO12] or low-level [CJO12] produced a relatively low performance. Besides, [GGS12] suggested that a combination of multimodal features, especially the metadata information, could improve hyperlinking performance.

MediaEval 2013

Table 2.1 compares the best submission of each group in terms of MAP[11].

[BPHPB13] created the linked segments based on provided transcripts and subtitles. The strategy was based on lexical information, computed from 20-word pseudo-sentence [BPHPB13]. Visual concepts were applied to re-rank the hyperlinking results. TF-IDF was used to index these segments and calculate the similarity. The fusion process used different fixed weights to textual and visual features, and the weights of the textual features (subtitles and transcripts) were

[10]http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

[11]MediaEval organisers have been investigating the evaluating mechanism of hyperlinking task since MediaEval 2012. In this thesis, all the results here are evaluated by the benchmarks developed after MediaEval 2014. Thus, the MAP values in Table 2.1 were lower than those published in papers. But the relatively ranks among participants were unchanged.

Table 2.1: A review of the best result of all participants in MediaEval 2013 hyperlinking task.

TEAM	MAP
Idiap2013 [BPHPB13]	0.5172
DCU [CJO13]	0.2354
LinkedTV13 [SHC+13]	0.2321
TOSCA-MP2013 [LSB13]	0.1887
UTwente [SAO13]	0.0609
soton-wais2013[PHS+13]	0.0594
HITSIRISA [GSGS13]	0.0474
MMLab [NNMdW13]	0.0376
UPC [VTAN13]	0.0240

higher. Table 2.1 showed that [BPHPB13] outperformed all other participants' runs in terms of MAP.

In [PHS+13], the author presented hyperlinking modes using textual information, including transcripts, synopsis, and video titles. Furthermore, they used SIFT descriptors to provide visual information. Experimental results demonstrated that visual information using SIFT tended to harm overall performance [PHS+13].

[VTAN13] utilised a similar strategy used in Video Google. They adopted SURF descriptors as visual features and K-means algorithm over SURF to build visual words. Experimental results showed that hyperlinking results using only visual features got lower MAP than those applying spoken transcripts.

[SAO13], [GSGS13], and [NNMdW13] used only transcripts to retrieve hyperlinks. In [GSGS13], transcripts were used for both anchors and videos, adopting the BM25 weighting strategy. A query video first was linked to the 50 most relevant videos, and each segment was detected based on lexical information. [NNMdW13] used time-based segmentation to create linked segments. Those segments were then enriched by extracting NEs using DBpedia Spotlight[12]. The Jaccard metric was applied to NEs to calculate the similarity between two enriched documents. [SAO13] used the posterior probability model to analyse the

[12]https://github.com/dbpedia-spotlight/dbpedia-spotlight

importance of query term, and expand the linked segments as the final results for evaluation. The results of [SAO13] were better than the other two groups. Besides, Table 2.1 demonstrated that their results were better than that in [VTAN13] involving only visual features (SURF)

In [LSB13], the content of linked segments was represented by three different multimodal features: ASR transcripts, metadata, and SIFT descriptors. The textual similarity was based on matching the spoken words and NEs from ASR transcripts or metadata. DBpedia Spotlight was used for the detection of useful words, and WordNet was used for query expansion. SIFT descriptors were used to re-rank the results. [LSB13] concluded that re-ranking strategy could provide small but consistent improvements.

The authors in [SHC+13] utilised a combination of visual concepts and spoken information. They determined the query content according to the transcripts and subtitles aligned at the corresponding query boundary. The hyperlinking results were re-ranked by the visual concepts. [CJO13] used transcripts, SIFT descriptors and metadata in its hyperlinking system. A fixed window in time was used to detect linked segments. TF-IDF algorithm was used to retrieve transcripts. Linear late fusion was adopted to combine multimodal features with equal weight for each feature. Both papers confirmed that integrating visual concepts and transcripts can improve the results.

Research works in MediaEval 2013 confirmed the importance of spoken transcripts. We conclude that the results using only transcripts achieve better MAP than that using only visual features. Furthermore, if combining transcripts and visual concepts can improve hyperlinking results. Besides, re-ranking the results using low-level features can slight increase MAP.

[BPHPB13] inspired us that using lexical information to detect potentially linked segment could work better than using a fixing window. In the remainder

Table 2.2: A review of the best result of all participants in MediaEval 2014 hyperlinking task.

TEAM	MAP
CUNI [GPKL14]	4.1824
LINKEDTV2014 [PMS+14]	0.2524
DCU [CJO14]	0.0791
JRS [BS14]	0.0556
IRISAKUL [SGSM14]	0.0335
DCLab [PFS14]	0.0135

of this thesis, we propose a set of experiments to compare the effectiveness between these two strategies.

[BPHPB13] used the combination of subtitles and ASR transcripts to represent query content and detect linked segments. Our investigation, however, will focus on only ASR transcripts since manually created subtitles could be unavailable in some other video collections.

MediaEval 2014

Table 2.2 shows the best result of all participants in terms of MAP in MediaEval 2014.

[PFS14] created the linked segments by cutting each video into shots according to the provided scene boundaries. The segment content was enriched with synonyms and conceptual terms from subtitles and transcripts detected by ConceptNet[13]. According to [PFS14], using manual subtitles can get better MAP.

[SGSM14] applied n-gram process to transcripts using Stanford Named Entity Recogniser, and used Latent Dirichlet Allocation (LDA) probabilistic topic models [BNJ03] to create a mixture of latent topics by indicating a probability distribution over n-grams.

[CJO14] applied word2vec with Wikipedia as the training collection to predict the potential context of named entities in transcripts. To detect linked segments,

[13]http://conceptnet5.media.mit.edu/

[CJO14] used sentence segmentations provided in transcripts. Experimental results show the ineffectiveness of using Wikipedia data to detect potential context.

[BS14] utilised textual information determined by a combination of subtitles, ASR transcripts and the metadata. VLAT [NPG13] on SIFT descriptors were used to match visual features [BS14] to re-ranking top results. Besides, [BS14] used the context around queries to enrich the query content. In conclusion, [BS14] confirmed the usefulness of context segments and the small contribution of SIFT descriptors.

[PMS+14] applied another query expansion strategy. They extracted the spoken terms from transcripts within the query segment. Then they used MoreLikeThis[14] and visual concepts to recreate the query content. [PMS+14] concluded that it is difficult to improve text based approaches when no visual cues are provided.

[GPKL14] utilised a fixed sliding window to create a baseline for hyperlinking. A segmentation employing decision tree was used to determine the segment boundary according to the lexicon information. Each segment was enriched by the context information extracted from the adjacent passages. The visual similarity was calculated by using Signature Quadratic Form Distance [BUS10]. Finally, the late fusion scheme was used in fusing visual and textual hyperlinking results where the fused weights were experimentally determined. According to [GPKL14], the best run demonstrated that lexical information and a combination of multimodal features were critical to achieving better hyperlinking performance. Besides, [GPKL14] demonstrated that when overlapping segments are preserved in the list, it could cause an overwhelming MAP (Table 2.2 illustrates that case), and concluded that hyperlinking results should filter overlapped linked segments

[14]https://cwiki.apache.org/confluence/display/solr/MoreLikeThis. MoreLikeThis constructs a lucene query based on terms within a document

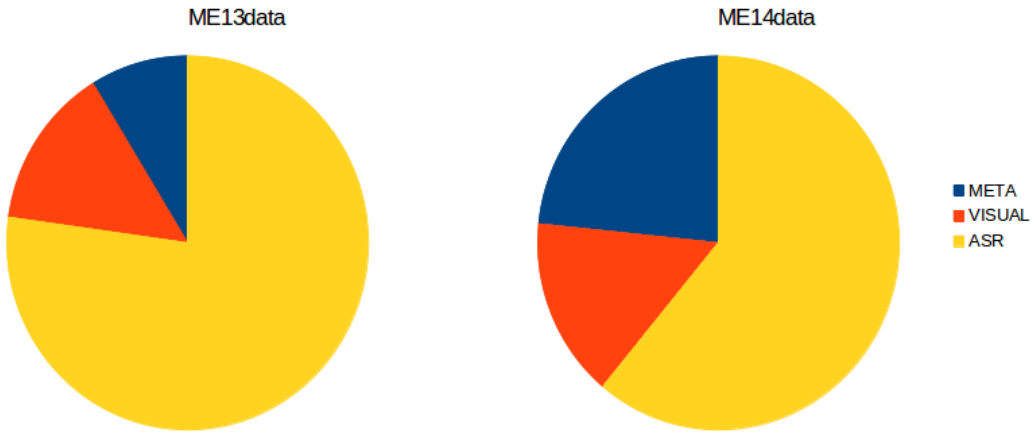


Figure 2.3: Multimodal features used for hyperlinking retrieval in both ME13data and ME14data collections

Research works in MediaEval 2014 revealed that combining the multimodal features is an effective approach to improve hyperlinking quality. [PMS+14] and [GPKL14] used visual features to improve the hyperlinking list retrieved by transcripts. Besides, research in [PMS+14] and [GPKL14] demonstrated the importance of recreating query content.

2.4.4 Discussion

Participants in MediaEval hyperlinking used multiple features to create video hyperlinks, including ASR transcripts, visual features, metadata, etc. We propose Figure 2.3 to demonstrate the statistics of multimodal feature usage in MediaEval Search and Hyperlinking task. In MediaEval 2013, 44 out of 57 submitted runs were constructed using ASR transcripts. Besides, 5 out of 57 used metadata features to determine hyperlinking similarity. The runs using visual features were only 5. In MediaEval 2014, the number of runs using visual features increased to 12. However, 47 out of 77 submitted runs still chose ASR transcripts as a retrieving feature. We concluded that ASR transcripts were widely used in MediaEval hyperlinking task. However, using only ASR transcripts can not achieve the best results. The review of research works in MediaEval concluded that using only

visual features, no matter low-level or high-level, got decreasing results compared with those of using ASR transcripts. Furthermore, research works in MediaEval 2013 [CJO14] and 2014 [GPKL14, PMS+14] pointed out that combining ASR transcripts with other features (high-level or low-level) could further improve hyperlinking performance. Besides, investigations in [GPKL14, PMS+14] showed that redefining query content can achieve better hyperlinking results. These conclusions inspire two primary research topics proposed in this thesis: data fusion to integrate multimodal features (discussed in Chapter 5) and query expansion (discussed in Chapter 6).

2.5 Chapter Conclusion

In this chapter, we reviewed multimodal features widely used in multimedia IR systems, data fusion techniques, and the development of multimedia hyperlinking. In Section 2.2, we presented a review of both low-level and high-level feature. In the remainder of this thesis, our content-based analysis for multimedia hyperlinking includes both multimodal features. To combine the hyperlinking results, we plan to utilise the data fusion techniques reviewed in Section 2.3. Section 2.4 reviews research works relevant to video hyperlinking. We reviewed the development of hyperlinking investigation in multimedia collection from early stage to MediaEval workshop. According to the review of MediaEval hyperlinking task, we conclude two state-of-the-art techniques in multimedia hyperlinking retrieval: integrating multimodal features and expanding the query content in hyperlinking.

In the next chapter, we will introduce the terminologies used in video hyperlinking, research questions to be addressed, and experiment hypothesis.

Chapter 3

A Multimedia Hyperlinking Framework

3.1 Chapter Overview

Last chapter reviewed papers on multimodal feature processing and multimedia hyperlinking construction. Existing research motivates our research hypothesis that using multimodal features could benefit video-based hyperlinking system. The remainder of this thesis discusses our investigation in video hyperlinking: Chapter 4, 5 and 6 are experimental chapters and describe how to use multimodal features to improve hyperlinking quality; Chapter 7 concludes the thesis; and the motivation of this chapter is introducing a set of high-level concepts to be used in experimental chapters. The content includes the structure of video hyperlinking framework in our experiment, the research questions to address, and the experiment hypothesis which applies to all experimental chapters.

This chapter consists of three sections. Section 3.2 introduces the architecture of our hyperlinking system. It outlines the purpose of the individual elements of our hyperlinking framework, including Query Anchors, Target Segments and Hyperlinks. Section 3.3 identifies the research questions to address in the experi-

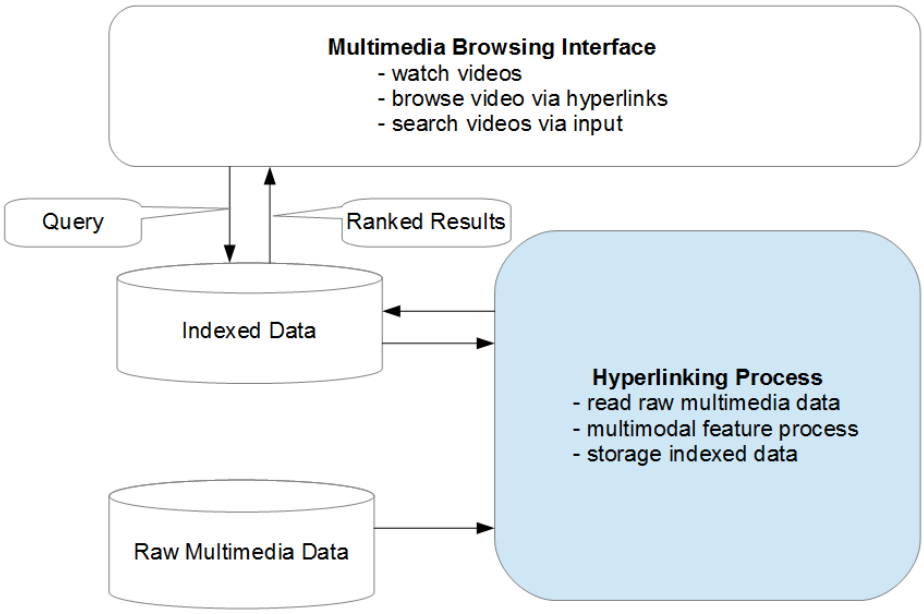


Figure 3.1: Multimedia hyperlinking system overview

mental chapters. We outline a total of 6 research questions and illustrate which chapter are going to address them. Section 3.4 presents the experimental hypothesis. It introduces the MediaEval data collections for hyperlinking task, describes evaluation benchmarks for multimedia hyperlinking according to [AEOJ13] and the workflow procedure for running crowdsourcing evaluation on the Amazon Mechanical Turk (AMT) website[1], and finally presents the design of experimental investigation.

3.2 Multimedia Hyperlinking System Overview

A multimedia hyperlinking system constructs hyperlinks within multimedia data collections. In general, multimedia data can involve different document types, such as formatted documents, still images, audio tracks, or video collections. Our research only focuses on video collections involving visual and audio information streams. Figure 3.1 shows a high-level overview of the hyperlinking system archi-

[1]https://www.mturk.com/

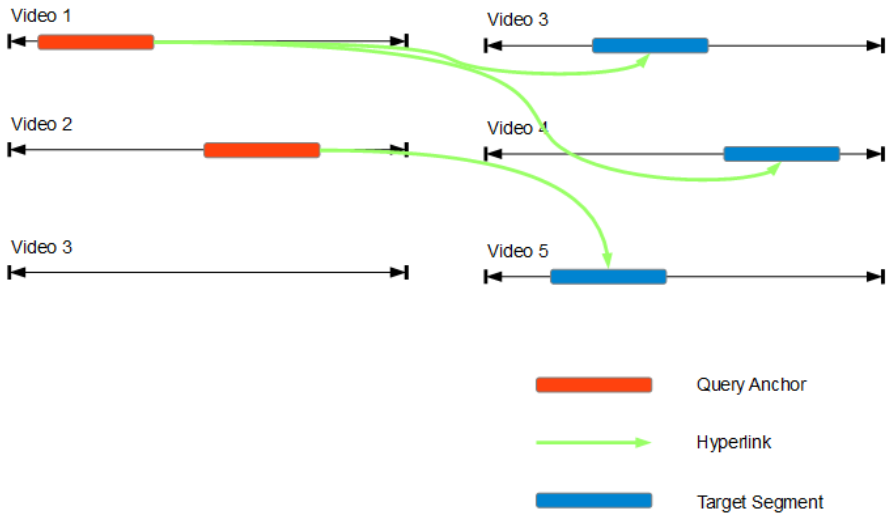


Figure 3.2: Hyperlinks between query anchors and target segments in a video collection

tecture that we investigate for our video archives. Similar to multimedia retrieval systems, a hyperlinking system presents an interface through which users can browse video segments associated with a set of ranked linked results. A hyperlinking process identifies all hyperlinked segments from within the multimedia collection.

The emphasis of our research is on the hyperlinking module, which is marked with a blue background in Figure 3.1. This module is responsible for processing raw multimedia resources to make them suitable for hyperlinking retrieval. It takes the multimedia raw data as an input, extracts the necessary multimodal features, and indexes them into a database. We define three essential elements in a multimedia hyperlinking system, as Query Anchor, Target Segment and Hyperlink. Figure 3.2 illustrates the role of these elements within the conceptual hyperlinking framework. The following sections introduce each element in our hyperlinking framework respectively.

3.2.1 Query Anchor

A query anchor is a clip of a video which simulate a user request to extend his/her browsing experience by linking to related content while watching a video. In an operational setting, a query anchor could be identified by the user as a region for which they wish to find relevant video segments, or it could be automatically identified. This assumes that users will be interested in some multimedia items presented in linked video clip. Items in the linked video could be people, objects, landmarks or spoken information. The definition of a query anchor shares some common points with the query input to an IR system. For the purpose of our investigation, a query anchor has the following features:

- A query anchor is well structured. Being a video clip, it is described according to two properties, the "jump in" point to indicate its start time in the source video and its duration to indicate its length from the start time.
- A query anchor description is typically composed of multimodal features. This consists of a combination of features extracted from the video from the "jump in" point to the end of this anchor.

We define a query anchor to be a video segment as a group of multimodal features which are representative in describing the details of the segment. A query anchor can be regarded as the query input for a hyperlinking system to identify relevant target video content. Since the query anchor only provides the "jump in" point and its duration, a hyperlinking system is free to determine the query content according to the multimodal features within it. In the remainder of this thesis, the experimental investigation explores how the determination of the query anchor content influences hyperlinking behaviour.

3.2.2 Target Segment

A target segment is a video clip within the collection that is assumed to be of interest to users navigating from a query anchor. A target segment is in some way semantically related to the query anchor to which it is linked. All or any multimodal features, including visual descriptors, audio content and transcripts, may be important in the formation of a hyperlink. The target segment shares some similar properties with retrieved documents in an IR system.

- The potential relevance of a target segment is measured by the similarity between the target segment and a query anchor as calculated using a matching function. State-of-the-art of similarity measures from IR are a potentially useful mechanism to determine this semantic similarity.
- The content of a target segment can be indexed and searched using traditional multimodal IR strategies.
- Semantic concepts in the target segment are described by the combination of multimodal features which can represent the users' interpretations when they are browsing the target segment.

However, [AEOJ13] notes out that a key difference between the video hyperlinking setting and traditional IR applications is that: the document units (target segments) for hyperlinking are not predefined, and that linking systems can thus return segments of arbitrary start point and length. Thus, a hyperlinking system needs to determine target segments from all the available video shots in a target video. We conclude our research consideration of target segments based on [AEOJ13]'s discussion in two aspects:

- Although can be arbitrary length, a target segment should be moderate in size. It essentially acts as a starting point from which users can explore the

content of the video to enrich their browsing experience. Users may lose patience when browsing an overlong video, while a short one could contain insufficient information.

- There is no obvious boundary for a potential target segments. Moreover, there are numerous methods to divide a target video stream into segments for matching, including shot boundaries, lexical information or the variance of visual features. It is impractical to create an infinite number of target segments to cover all the potential interesting points for linking. Therefore, an algorithm is required to efficiently identify those segments with a higher priority of being useful for hyperlinking.

.

To construct content-related hyperlinks across a multimedia collection, a hyperlinking system should extract potential target segments, index their content for searching, and retrieve potentially relevant segments according to query anchors. Therefore, methods for identifying potentially relevant target segments will be the first research issue to address in the remainder of this thesis.

3.2.3 Hyperlink

A hyperlink is a connection between a query anchor and a target segment, indicating a semantic relationship between them. Each query anchor can create multiple hyperlinks to target segments which will have different strength of allocation. Each target segment can also be linked to multiple query anchors. Figure 3.2 shows an overview of hyperlink type in a multimedia collection. It should be noted that a hyperlink can be categorised as either within an individual video and within collection. Figure 3.3 shows an example of within document and within collection hyperlinks. Linking within a video is likely to direct the user to

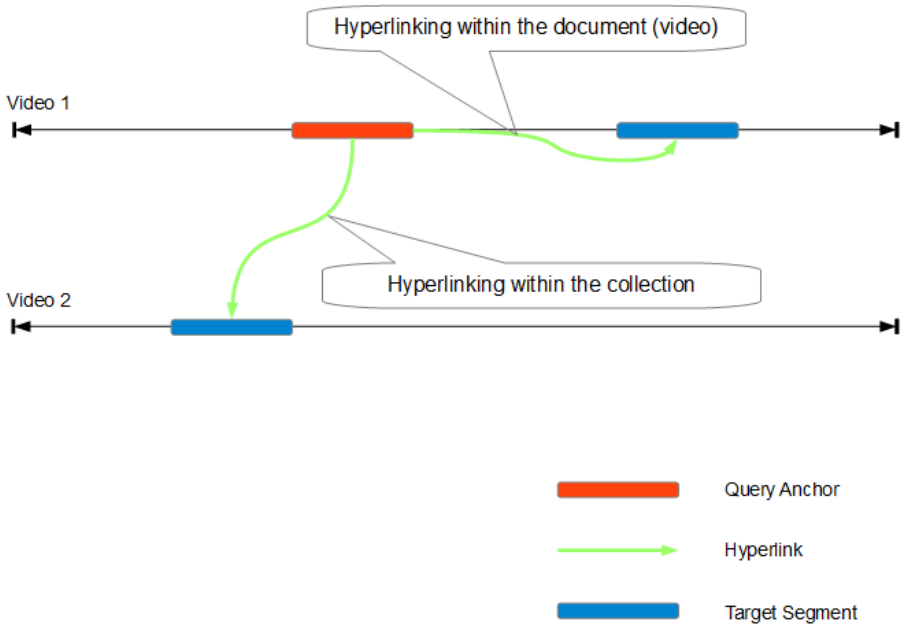


Figure 3.3: Within document and within collection hyperlink targets

a segment which is strongly related to the source, while links to other videos are likely to have more diverse targets. In our experimental investigation, we take into account both types of hyperlinks across the test collection.

3.2.4 Discussion

This section introduces the elements constructing the hyperlinking system used in the following experimental chapters. Our investigation is dedicated to addressing research issues of applying multimodal features to build query anchors, determine target segments, and finally create hyperlinks. In the following section, we introduce the details of research questions relevant to these elements in each experimental chapter.

3.3 Research Questions

We propose our research questions by combining the research objectives proposed in Chapter 1 and the details of video hyperlinking system introduced in the previous section. In the remainder of this thesis, the experimental chapters (Chapter 4 to Chapter 6) will focus on 6 research questions (RQ):

- RQ 1: How do classic IR models and textual features benefit hyperlinking retrieval?
- RQ 2: Can we efficiently identify target segments in terms of improving hyperlinking retrieval quality?
- RQ 3: How do other multimodal features except textual influence hyperlinking retrieval?
- RQ 4: Can we improve data fusion strategies to integrate multimodal features for both ME13data and ME14data?
- RQ 5: How does recreating query anchor content improve hyperlinking retrieval?
- RQ 6: Can integrating query anchor recreation and multimodal features further improve hyperlinking results?

Figure 3.4 illustrates the structure of our hyperlinking framework and RQs to be investigated in each chapter. Experimental investigations use the data collections, ME13data and ME14data, introduced in Section 3.4.1. RQ 1 and 2 are supposed to be investigated in Chapter 4. Experiments will focus on how to use textual features to define target segments and retrieve hyperlinks. The conclusion of Chapter 4 will be used as the baseline for the remainder of experimental chapters. Investigation on RQ 3 and 4 will be proposed in Chapter 5, and our

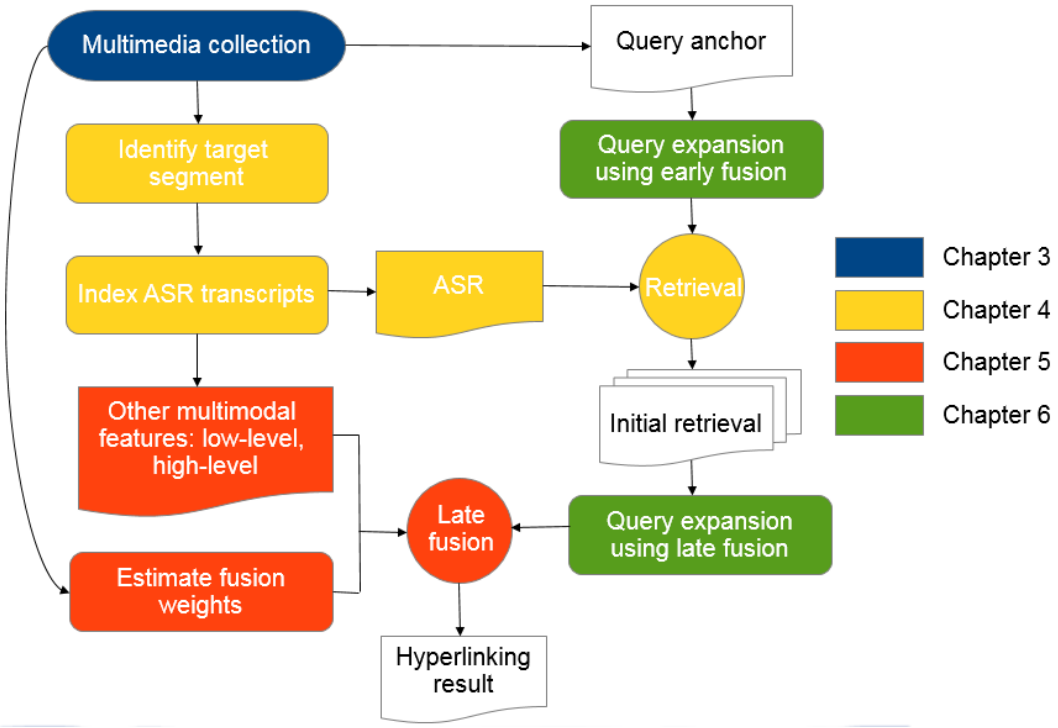


Figure 3.4: Introduction of research questions to address in each experimental chapter

research will involve other multimodal features and demonstrate how to integrate them to improve hyperlinking performance. Chapter 6 will investigate the issue of recreating hyperlinking queries. Experiments demonstrate that using expanded queries can further improve hyperlinking results concluded in Chapter 4 and 5.

To address these research questions, we design a set of experiments to demonstrate the effectiveness of different multimodal features in the remainder of this thesis. Before experimental chapters, it is necessary to introduce general principles. This is our motivation to propose next section.

3.4 Experimental Hypothesis

We define a set of concepts, including the data collections used in the future experiments, the evaluation benchmarks to determine hyperlinking quality, and the workflow of experimental investigation, as the experimental hypothesis. All

experiments and discussions in the following chapters involve these concepts to address our research questions. Thus, we introduce these concepts in this section so that readers can have a global view of our experimental design.

3.4.1 Multimedia Data Collection

The multimedia data collection used for the hyperlinking study described in this thesis originates from the Search and Hyperlinking tasks at MediaEval 2013 and MediaEval 2014. We use the abbreviation **ME13data** and **ME14data** to respectively denote these data collections. The motivation of selecting these two collections are:

- The data collections in MediaEval 2013 and MediaEval 2014 are provided by BBC company. They contain rich multimedia contents extracted from BBC online TV plays. We can examine our research ideas in the collections used in the real world.
- Multiple research groups have contributed the development of multimodal features in these collections. In this thesis, we bring some existing research achievements to board our investigation in multimodal features.
- AXES[2] groups were dedicated to investigating these collections and provided funding support to create the hyperlinking ground-truth (introduced in 3.4.4) using crowdsourcing. Using the corresponding ground-truth in these collections, we can examine our approaches in terms of user judgement.

MediaEval 2013 Data Collection (ME13data)

The ME13data contains 1,260 hours of TV video provided by the BBC. It contains broadcast content between 01.04.2008 to 11.05.2008. The videos in the ME13data

[2]http://www.axes-project.eu/

collection involve various types of TV shows including BBC News reports, TV drama, documentary films, entertainments, etc. The average length of a video is roughly 30 minutes and all videos are in the English language. The total number of videos is 2,323. In this thesis, we propose to use a total of 21[3] valid queries defined by the MediaEval 2013 workshop.

MediaEval 2014 Data Collection (ME14data)

ME14data is a collection of over 2,686 hours of TV video provided by the BBC. The content collection was originally broadcast between 12.05.2008 and 31.07.2008. ME14data consists of the same types of TV programme as ME13data. The average length of a video in ME14data is roughly 45 minutes and all videos are in the English language. The total number of videos in ME14data is 3,520. The MediaEval 2014 workshop provided 30 query anchors associated with the corresponding ground truth.

Multimodal Features in ME13data and ME14data

The MediaEval task organisers provided various sets of multimodal features for both ME13data and ME14data. We introduce the features which are used in our experimental investigations described in this thesis.

- **Spoken Transcripts** Two automatic speech recognition (ASR) transcripts were created for the ME13data and ME14data by LIMSI/Vocapia Research[4] and LIUM Research team[5]. The implementation of spoken transcripts from LIMSI/Vocapia was based on the method described in [LG08]. The LIUM system is based on the CMU Sphinx project [RBD+11]. The LIUM algorithm

[3]The MediaEval 2013 defined 98 query anchors and provided the corresponding ground-truth for 30 of them. We select 21 out of these 30 queries by removing those located within the same video scenes.

[4]http://www.vocapia.com/

[5]http://www-lium.univ-lemans.fr/en/content/language-and-speech-technology-lst

builds the transcripts associated with spoken words and the corresponding timestamp. The LIMSI algorithm provides more spoken information including not only spoken words and their time stamps in the video stream, but also speaker identification and the segmentation of spoken data in terms of sentences. In the remainder of this thesis, we use the abbreviations of LIUM and LIMSI to represent the corresponding transcripts respectively.

- **Visual Features** ME13data and ME14data contain a set of visual descriptors of the video content, including automatically detected shot boundaries, one automatically extracted keyframe per shot, and the outputs of concept detectors. For each video, shot boundaries are determined and a single key frame per shot is extracted using a system kindly provided by Technicolor [MLD+06]. In total, the system extracted approximately 1,200,000 shots/keyframes for ME13data, and 1,500,000 shots/keyframes for ME14data. In Chapter 5, we give a detailed description of the high-level concepts used in our experimental investigation.

- **Video Metadata** Each video in ME13data and ME14data has associated metadata, which was manually created by BBC. The metadata includes a set of textual attributes including "video title", "uploading date", "description", "uploading author", etc. The "description" is the primary attribute used for the further experimental analysis in the remainder of this thesis. In Chapter 5, a detailed description of the experimental design is introduced.

In the remainder of this thesis, we use the word "multimodal features" to represent all these three multimedia features (spoken transcripts, visual features and video metadata) in the ME13data and ME14data. Moreover, according to Section 2.2, we regard the spoken transcripts as the textual low-level feature since spoken transcripts directly represent the word presented in the video stream rather than cognitive concepts built by knowledge and experience from human activities.

The video metadata that provides an overall description of the corresponding video content is regarded as the high-level feature. The visual features involve both low-level and high-level types whose details are introduced in Chapter 5.

3.4.2 Ground Truth Construction via Crowdsourcing

Multimedia hyperlinking evaluation for research purposes is based on evaluating the quality of each participant's hyperlinking creation runs in terms of various metrics. In order to evaluate these metrics, a ground truth is required to indicate whether the target segment of a proposed hyperlink is relevant to the current query anchor or not. In this section, we[6] introduce the mechanisms used to construct the hyperlinking ground truth utilising the online human resources.

Firstly, it is important to realise that relevance as judged in an IR system is a personal assessment [BYRN99]. The relevance between query anchors and target segments discussed in this thesis is determined in terms of human perspective, which follows the hypothesis introduced in Search and Hyperlinking tasks in MediaEval 2013 and 2014. Thus, to build the ground-truth, we need human judgement to identify the relatedness of retrieved hyperlinks. In Chapter 2, we described how human annotation was used in the NTCIR CLLD task to determine the relevance of linked Wikipedia documents. A similar strategy was applied to evaluate retrieved hyperlinks in the MediaEval hyperlinking task. Crowdsourcing hires online workers to carry out well specified large scale human centered tasks. In this process, a requestor publishes a task and recruits online workers to carry it out, creating a Human Intelligence Task (HIT) which describes the task. A crowdsourcing platform implements HITs to satisfy the requestor's requirement and specifies the reward which is available for those workers completing the HITs [Jon13]. In our work, a HIT allows researchers to

[6]The author of this thesis has participated in constructing the ground truth of ME13data and ME14data. The author's primary duty was designing the crowdsourcing user interface, collecting user feedback from the crowdsourcing website, and examining the validation of these questions.

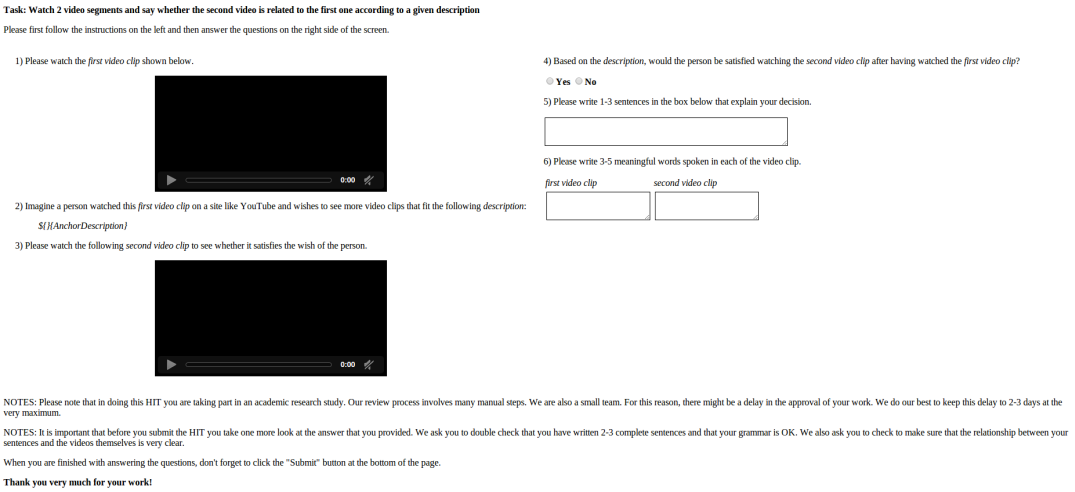


Figure 3.5: The screenshot of Amazon Mechanic Turk (AMT) assignment

obtain human-generated feedback about the relatedness between video segments, i.e. whether the hyperlinks that we proposed are potentially valuable for real users. The construction of a ground truth for hyperlinking retrieval uses the crowdsourcing platform to identify targets related to the video anchors.

In MediaEval 2013 and 2014 Search and Hyperlinking tasks, the hyperlinking ground truth was created using the Amazon Mechanical Turk (AMT)[7] crowdsourcing platform. After collecting the participants' submission, a set of video pairs involving each query anchor and the linked segments was constructed. Each video pair was uploaded to the AMT platform. The AMT platform presented the assigned workers with assignments containing the linked segment associated with the corresponding query anchor. The AMT workers were required to watch the two video segments and answer a number of questions to indicate the level of relevance of the target to the query and to explain their decision. Figure 3.5 shows a screenshot of the crowdsourcing task. Each AMT worker was required to answer three questions as follows:

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

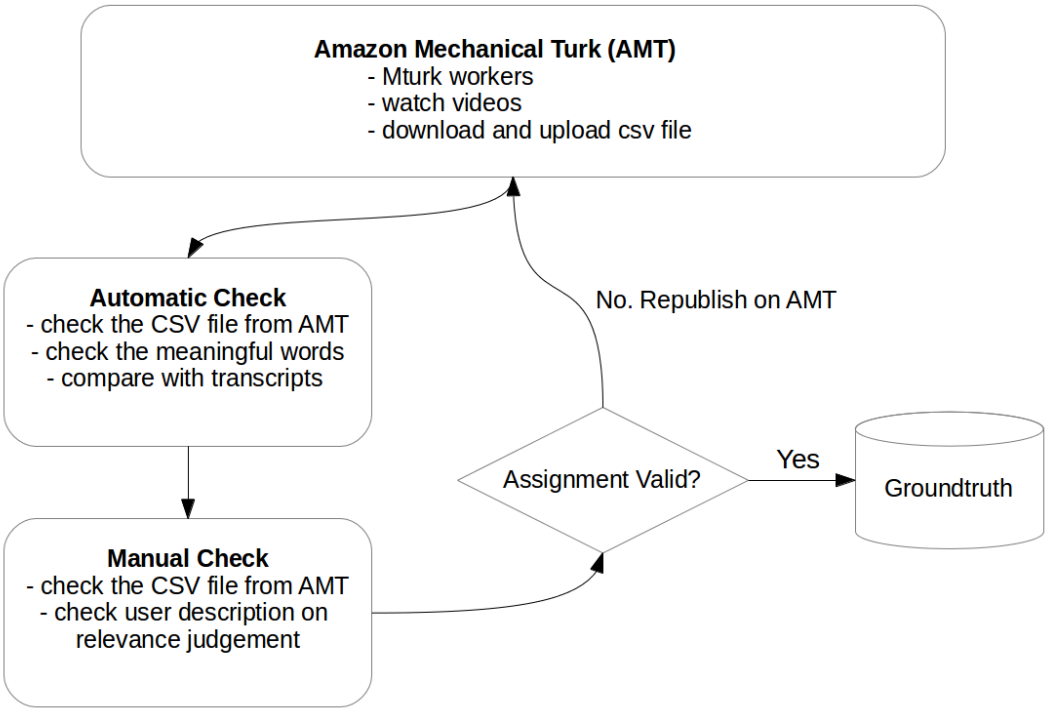


Figure 3.6: Crowdsourcing flow to annotate groundtruth

- Question 1. Based on the description, would the person be satisfied watching the second video clip after having watched the first video clip?
- Question 2. Please write 1-3 sentences in the box below that explain your decision.
- Question 3. Please write 3-5 meaningful words spoken in each of the video clips.

The purpose of these questions was to collect user feedback on the relevance judgement and to enable validation of the assignment input. The first question was used to collect the AMT workers' judgements on the relatedness of the video pair. The second question was used to check whether the crowdsourcing workers watched the video content properly. The third question was used to validate the first question by describing the reason for making this judgement. Figure 3.6 shows the workflow of the crowdsourcing procedure. Each crowdsourcing

worker was required to complete all three questions. A crowdsourcing assignment was regarded to be invalid if:

- The assignment for AMT workers was incomplete. Automatic checking was used to scan all the results returned from the AMT website in the format of CSV files and filter out those assignments with empty answers.

- The assignment worker provided incorrect meaningful words spoken in the presented video clips. A script was developed by the hyperlinking task organisers to index all spoken transcripts contained in the corresponding clips. All the answers were examined automatically by this script. If an error description was caused due to spelling errors, an unrecognised input format or an informal expression, it was accepted as valid. In all other cases, the assignment was rejected. The valid answers were collected for a further manual validation to determine the reason for a user's judgement on the relevance of the hyperlinking.

- The reason for worker's judgement on the relevance of the video is not persuasive. We expect crowdsourcing workers to make a reasonable decision when determining the hyperlinking relevance. In this stage, manual validation was applied to check the answers of the third question. A reasonable answer that could be accepted such as "the two segments are about music", "they are both BBC breakfast News", or "the same reporter interviews different people". Answers providing a fuzzy description, like "I love this video", "The music is quite wonderful" were regarded as invalid.

We politely rejected invalid crowdsourcing works, and republished these questions on AMT until we had collected valid answers for all video pairs.

After collecting all the valid answers, we analysed the relatedness of each video pair according to the first question. We categorised the answers into positive and negative assignments. A video pair in the ground truth is positive if the

crowdsourcing worker regards the linked video as related to the query anchor. If not, the video pair is negative. Each video pair was judged by two crowdsourcing workers whose answers were both validated. In this thesis, we apply the strategy used in MediaEval 2013 and MediaEval 2014 that a video pair is regarded to be negative only if neither of the crowdsourcing workers indicates them as relevant. In Appendix A (Tables A.1 and A.2), we show the comparison of the two Mturk workers' judgements on the relevance between a query anchor and the corresponding segment in the ground truth pool. The ground truth contained a total of 22,313 validated video pairs. In the ME13data, crowdsourcing evaluated 21 queries and collected 9,973 video pairs composed of 2,982 positive and 6,991 negative ones. In the ME14data, crowdsourcing evaluated 30 queries and collected 12,340 video pairs composed of 1,888 positive and 10,452 negative ones.

3.4.3 Evaluation Metrics

An ideal hyperlinking framework should be able to construct hyperlinks to relevant target segments without linking to non-relevant ones. Furthermore, it should at least assign higher rank to relevant segments than non-relevant ones. Thus, similar to state-of-the-art IR evaluation, evaluating hyperlinking performance is based on the measures of relevance in the retrieved ranked results. Two primary metrics are utilised to evaluate the recall and precision rates for the further hyperlinking experiments in this thesis according to [AEOJ13].

Evaluation Metrics

The main metrics in this thesis evaluate hyperlinking performance in terms of the precision rate in the top N linked result (P@N) and the Mean Average Precision (MAP) value. Equation 3.1 defines P@N which takes all retrieved target segments at a given cut-off rank N into account,

$$\text{P@N} = \frac{|\text{relevant target segments@}N|}{N}, \tag{3.1}$$

where |relevant target segments @ N| means the number of linked target segments relevant to the current topic in the top N results. Equation 3.2 defines the MAP metric, which computes the average of the precision value over all the relevant ranked items in the retrieval list,

$$\text{MAP} = \frac{\sum_{q=1}^{Q} \text{AveragePrecision}(q)}{|Q|}, \tag{3.2}$$

where $|Q|$ is the number of queries, and the function AveragePrecision for a query q is defined as follows:

$$\text{AveragePrecsion} = \frac{\sum_{i=1}^{N} P@i}{|\text{relevant target segments}|}, \tag{3.3}$$

where $P@i$ returns the precision rate at the rank i, and |relevant target segments| returns the total number of relevant segments in the ground truth pool. The previous section introduced the methodology used to determine the relevant segments for each query anchor by using the AMT platform. However, it is impractical to evaluate all potentially linked segments due to the significant cost involved. Therefore, the evaluation metrics need a specific mechanism to determine the relevance of a linked segment associated with the ground truth where there is a partial overlap between the linked segment and the relevant content. In the remainder of this section, we introduce various mechanisms developed within the MediaEval Search and Hyperlinking task to attend to compare the effectiveness of the submitted results.

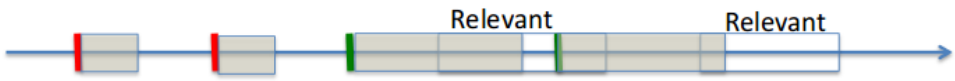


Figure 3.7: Overlap evaluation, source: [AEOJ13]

Overlap Mechanism

The overlap mechanism defines that a target segment is considered as relevant if it overlaps with a relevant segment [AEOJ13]. This identifies a linked segment as relevant if it overlaps a video segment indicated as positive in the ground-truth. Figure 3.7 shows an example of overlap evaluation in which the linked segments starting with the green "jump in" point are regarded as relevant due to the overlap with the relevant ground truth. In our experiments, we use the abbreviations P@N and MAP to represent the evaluation metrics for the overlap mechanism.

Bin Mechanism

The overlap mechanism doesn't take into account the context information of positive segments in the ground-truth. Therefore, [AEOJ13] also introduced the bin mechanism. This assigns the result segments to units of a fixed size, which are referred to as "bins" [AEOJ13]. The video stream is divided into a set of bins of equal duration. The duration of each bin is set to be 120 seconds in [AEOJ13] following the maximum size of linked segment defined in the MediaEval Search and Hyperlinking task [EJC+13]. Each bin is defined as relevant if there is at least one relevant linked segment located within it. Finally, the ground truth consists of relevant and irrelevant bins to evaluate the hyperlinking results. If a linked segment overlaps with a relevant bin, it is regarded as relevant to the current topic. Figure 3.8 shows an example of bin evaluation.

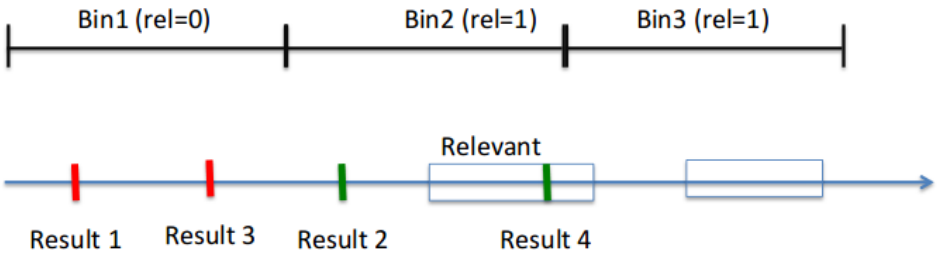


Figure 3.8: Bin evaluation, source: [AEOJ13]

Tolerance of Irrelevance Mechanism

The bin mechanism takes context information into account to improve hyperlinking evaluation quality. A potential risk of the overlap and bin mechanisms, however, is that a large number of short segments may be concentrated on a particular video shot due to the sharing of similar context information. A group of segments concentrated within the same positive bin achieves a higher precision and MAP value. However, this situation has little contribution to improving the user's browsing experience. The tolerance of irrelevance mechanism modifies the relevance judgement in terms of the following principles. Each relevant ground truth is associated with a tolerance area with respect to context information. A retrieved target segment, if its start time is located within the tolerance area, is set as relevant. Each tolerance area is only encountered once. When evaluating a ranked hyperlinking result, the algorithm traverses it from the top result. If a relevant bin overlaps a linked segment, the algorithm regards this segment as relevant. Any linked segment overlapping this relevant bin at a lower rank won't improve the corresponding benchmarks since this bin has been "seen" by a linked segment at a higher rank.

Figure 3.9 shows an example of the tolerance of irrelevance mechanism. It illustrates that Result 2 is judged as positive since its start time is encountered

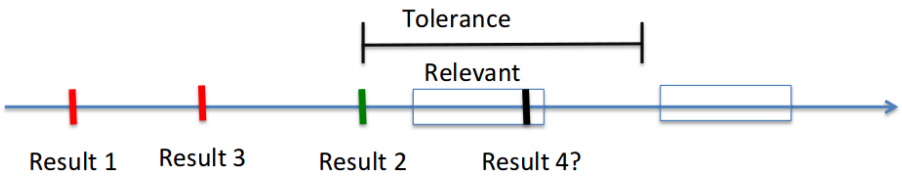


Figure 3.9: Tolerance of irrelevance evaluation, source: [AEOJ13]

by the tolerance area. While Result 4, even located in the tolerance area, is not counted as positive since this area has been seen by Result 2.

In conclusion, overlap evaluation is a metric developed within in the MediaEval 2012 Search and Hyperlinking task to MediaEval 2014. While the other two, bin evaluation and tolerance of irrelevance evaluation, were proposed for MediaEval 2014 Hyperlinking task to complement the overlap metric. In this thesis, the primary metrics to examine our experimental results are the overlap metrics (MAP and P@N). Furthermore, tolerance of irrelevance evaluation (tMAP) performs as a complement to overlap evaluation (MAP) in our experiments.

3.4.4 Experimental Design

This section introduces the principles which are applied for all the experiments described in experimental chapters. Figure 3.10 illustrates the workflow to retrieve segments for creation of hyperlinks for each query anchor. We discuss four concepts which are applied in the future experimental chapters: data indexing, retrieval results, score normalisation, and filter overlapped results.

- **Data Indexing**: All multimedia raw data is indexed and searched according to proposed IR models. We use Apache Lucene 4.9.0[8] software in order to index and retrieve the target segments with spoken information. A standard analyser component of Apache Lucene is used to convert text data into

[8]http://lucene.apache.org/core/

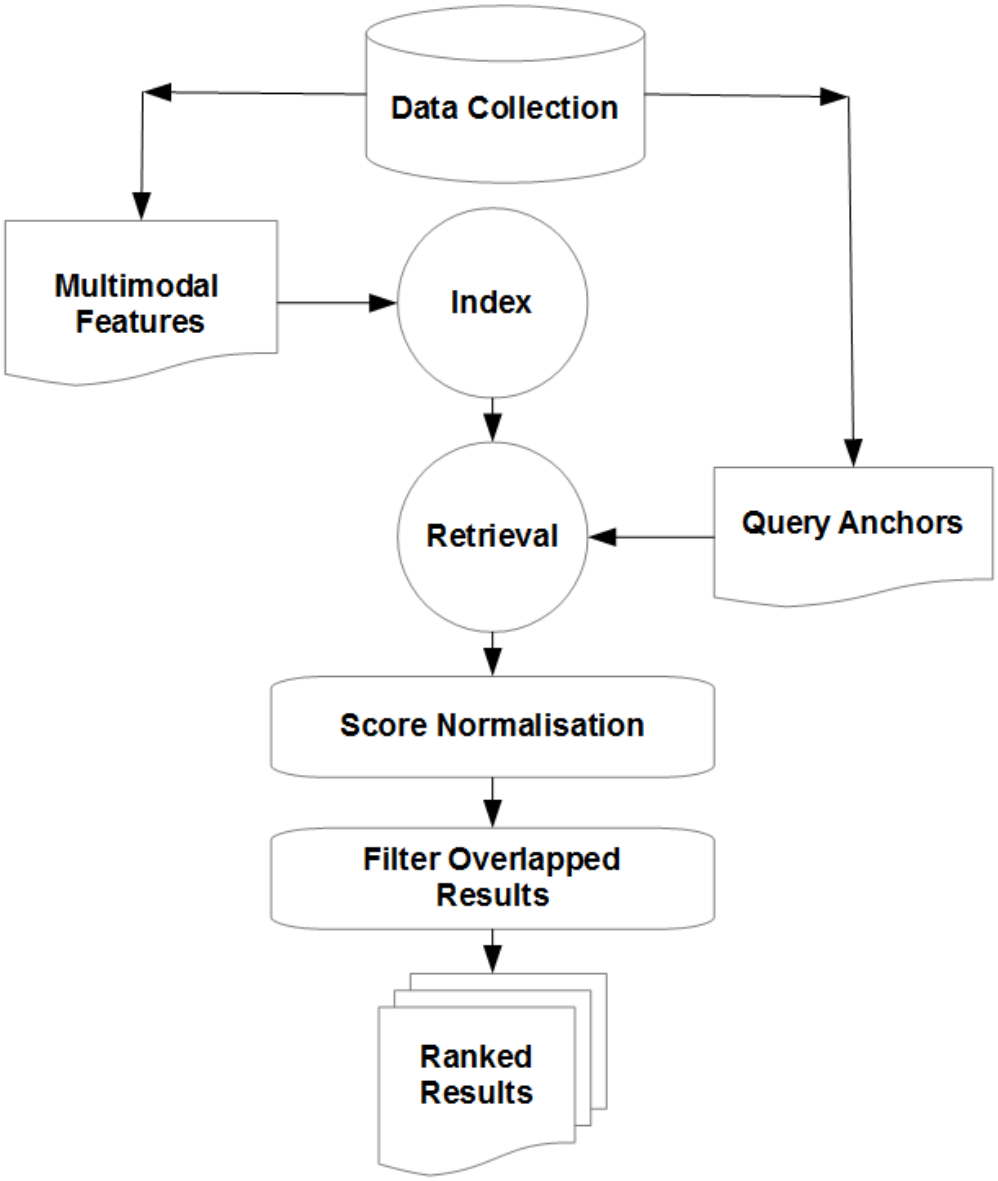


Figure 3.10: Experiment hypothesis: the workflow of the research hyperlinking system

the format for search within Lucene. Each spoken word is converted into lower case. The stop words are removed using the default list provided from Lucene. The analyser tokenises text-based using sophisticated set of grammar rules which recognises e-mail addresses, acronyms, and alphanumerics characters [Son09]. We use Porter Stemming [Por80] algorithm to implement word stemmer.

- **Retrieval Results**: The retrieval results of each experiment involves multiple runs. Each run consists of N linked target segments which are ranked by their scores. The value of N is experimentally set to 1,000 following a general principle defined in Search and Hyperlinking task of MediaEval 2014.

$$s_{nor}(r) = \frac{s(r) - \text{MinScore}(R)}{\text{MaxScore}(R) - \text{MinScore}(R)} \tag{3.4}$$

- **Score Normalisation**: The normalised scores are used in the multimodal fusion proposed in Chapter 5 and Chapter 6. For each retrieved result scored $s(r)$ at rank r, the normalised score is calculated using Equation 3.4, where MinScore(R) and MaxScore(R) output the minimum and maximum scores in the ranked list R respectively. Score normalisation has no influence on the relative position of retrieved segments.

- **Filtering Overlap**: The target segment identification algorithm indicates potential video segments to be retrieved. It follows a simple principle that detected segments should cover the most multimodal information. Therefore, extracted potential segments could share an overlap with others. In Chapter 2, [GP13] concluded that containing overlapped segments in hyperlinking results could overwhelm MAP score. Thus, our hyperlinking process applies filtering steps to remove the overlapped segments in the ranked list: if a set of retrieved segments share an overlap, the hyperlinking system keeps the one with the highest rank and removes all others.

In our experimental system, data indexing happens after collecting multimodal features and before retrieving hyperlinking results. After achieving all retrieved results, the hyperlinking process applies score normalisation, filters the overlapped results, and generates the final results for evaluation and discussion. This workflow applies the whole experimental investigation.

3.5 Chapter Conclusion

This chapter presented a set of high-level concepts to be used in experimental chapters. We introduced three essential elements to index and search multimedia hyperlinks from raw multimedia resources: query anchor, hyperlink and target segment, in Section 3.2. Then, in Section 3.3, we proposed the research questions to address in our investigation. Section 3.4 outlined our experimental hypothesis. This section involves the description of test data collections, the evaluation benchmarks and how to build the ground-truth using AMT, and our experimental design. The contribution of this chapter is presenting a high-level view of our hyperlinking system. Readers can understand the workflow of our hyperlinking investigation and the corresponding concepts used in the future experimental chapters.

The following chapters (Chapter 4, 5, and 6) are our experimental chapters. We investigate the performance of multimodal features in the proposed hyperlinking system. However, multimodal features involved in a multimedia collection present additional challenges. Therefore, we decide to use a simplified multimedia hyperlinking framework to investigate how to address these issues. It involves only text information from multimedia data collection without further processing on visual cues for the main reason that research on text-based IR is well-developed and thus forms a good baseline. Using a text-based data collection, we plan to investigate different strategies to identify target segments and evaluate hyperlinking models using multiple benchmarks.

Chapter 4

Creating Hyperlinks using Transcripts of Spoken Data

4.1 Chapter Overview

In Chapter 2, we introduced that ASR transcripts are widely used in video-based retrieval or hyperlinking and concluded that combining ASR transcripts with other multimodal features can improve hyperlinking retrieval. The motivation of this chapter is discussing how ASR transcripts influence hyperlinking retrieval individually. We are dedicated to solving two research questions (RQ):

- RQ 1: How do classic IR models and textual features benefit hyperlinking retrieval?
- RQ 2: Can we efficiently identify target segments in terms of improving hyperlinking retrieval quality?

Reviews in Chapter 2 outlined some conclusions in MediaEval 2013 and 2014 respectively, and those conclusions inspire our strategies of hyperlinking in ASR transcripts. In the remainder of this chapter, we propose: ASR transcripts indexing and retrieval using different weight models, target segmentation identification,

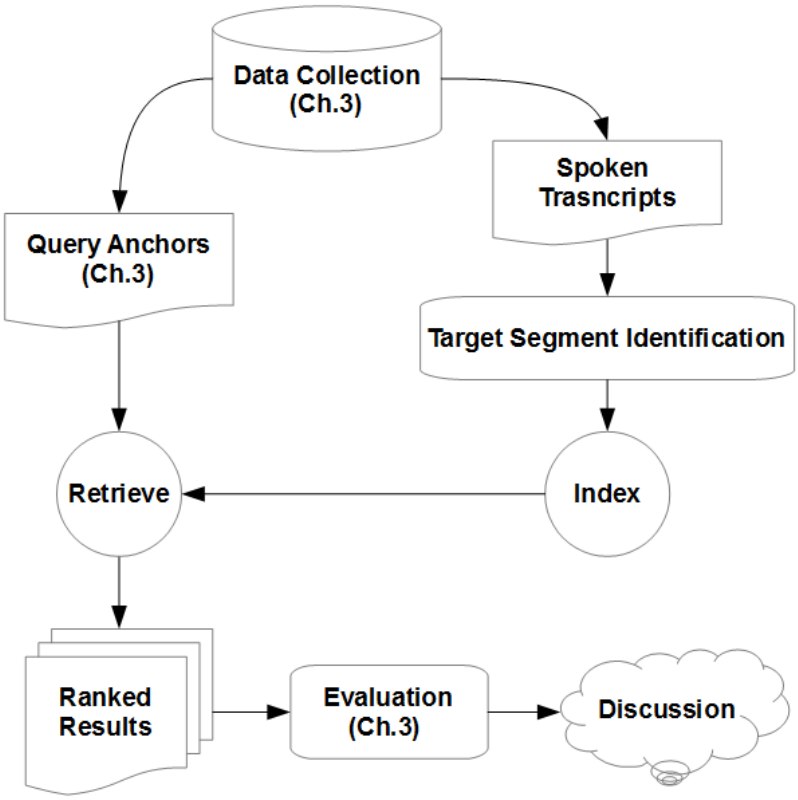


Figure 4.1: An overview of research design

and parameter setting for term weight models. We will not only examine which approach can achieve better hyperlinking in terms of our evaluation benchmark introduced in Chapter 3, but also investigate whether those approaches can reach an agreement on the parameter setting (or not) in both data collections. The results of this investigation provide evaluating baselines for the remainder of the thesis.

This chapter is structured as follows. Section 4.2 briefly introduces two classic text IR models which are applied for hyperlinking using the spoken data. Experimental results compare their hyperlinking performance for various ASR transcripts. Section 4.3 investigates different strategies to identify target segments. The hyperlinking framework is constructed according to the optimal text IR model and ASR transcripts introduced in Chapter 3.1. Experimental results show how

Table 4.1: Acronyms in experimental investigation

Abbreviation	Description
ASR	The spoken transcripts detected by automatic speech recognition
TF-IDF	TF-IDF algorithm to index and retrieve ASR
BM25	Okapi BM25 algorithm to index and retrieve ASR
TSW	Use the time-based sliding window
CSW	Use the content-based sliding window
W	Detect the segment boundary in terms of the spoken words
S	Detect the segment boundary in terms of the spoken sentence

different algorithms of identifying target segments influence hyperlinking performance. Section 4.4 investigates the mechanisms of determining appropriate parameters to index and retrieve transcripts using Okapi BM25 model. Section 4.5 concludes the chapter.

For readers' convenience, we propose Figure 4.1 and Table 4.1. Figure 4.1 illustrates a global view of experimental design proposed in this chapter[1], and Table 4.1 shows the acronyms of multimodal features and research methodologies. Reader can understand our experimental design proposed in this chapter, and have a reference to check the acronyms in each experimental discussion.

4.2 Spoken Information Retrieval Model

In this section, we present an overview of two text IR algorithms which we apply for the task of multimedia hyperlink creation based on ASR transcripts. First, we introduce the classic vector space model, and then the BM25 probabilistic model.

[1]We also mark those research terms already introduced in previous chapters with the abbreviation "Ch.". For example, a part marked with Ch.3 means that this term has been discussed in Chapter 3.

4.2.1 Vector Space Model

Vector Space Model (VSM) is a classic IR model in which each document is represented as a vector of identifiers [SWY75]. The definition of an identifier could be a single word or an n-gram, depending on different retrieval systems. In this thesis, we describe each identifier as a term. Each document D_i in the data collection is represented as a vector $D_i = [w_{1,i}, w_{2,i}, ...w_{k,i}]$. The dimensionality of a vector is the size of term vocabulary. Each weight $w_{k,i}$ is a term defined based on different IR models. If any term occurs in a document, the corresponding weight is set to be non-zero. The cosine distance algorithm is widely applied to calculate the similarity between two document vectors as:

$$\text{sim}(d_i, d_j) = \frac{\sum_{k=1}^{N} w_{i,k} \cdot w_{j,k}}{\sqrt{\sum_{k=1}^{N} w_{i,k}} \cdot \sqrt{\sum_{k=1}^{N} w_{j,k}}}, \tag{4.1}$$

where N is the number of terms in a document vector, and $w_{i,k}$ and $w_{j,k}$ represent the weight of ith/jth term in the document D_i/D_j respectively. There are multiple term weighting strategies to determine the value of w. Among them, a simple strategy is the Boolean Retrieval Model which sets a term weight w to be either 0 or 1, meaning a word exists in a document or not. However, using boolean model can not show the diversity of a term existing in different documents. An alternative solution is calculating term weights according to the apperance of a term within a document collection. In the following section, we briefly review a classic term weighting mechanism known as *TF-IDF*.

TF-IDF Term Weighting

TF-IDF term weighting, also known as term frequency-inverse document frequency weighting, is a classic term weighting model. The weight w of each term t is determined by its own term frequency $tf(t, d)$ in a document d and its inverse

document frequency $idf(t, d, D)$ within the search collection. The definition of term weight $w_{t,d}$ is shown as following:

$$w_{t,d} = tf(t, d) \cdot idf(t, d, D), \tag{4.2}$$

$$idf(t, d, D) = \log \frac{|D|}{|\{d \in D | t \in d\}|}, \tag{4.3}$$

where D is the total number of documents in the document collection.

Multimedia hyperlinking in this chapter regards each single spoken word presented in an ASR algorithm as a term in the VSM model. *TF-IDF* is applied to calculate the weight of each term. Each target segment is regarded as a potentially relevant document for retrieval. Therefore, the term frequency is determined according to the appearance of a spoken word in a target segment. The inverse document frequency is calculated according to the number of target segments in which a spoken word exists and the total number of potential target segments within the multimedia collection.

4.2.2 Probabilistic Retrieval Model

The basic principle of the probabilistic retrieval model is to estimate the probability that a document d is relevant to a query q [RvRP81]. Generally, it is assumed that a document is either relevant or non-relevant to the query. The relevant set is defined as R and all other documents are designated as $\overline{R}$. The probabilistic retrieval model determines the similarity by maximising the overall probability defined in Equation 4.4. The most well known practical instantiation of the probabilistic retrieval model is the Okapi BM25 weighting model (BM25) [RWJ+95].

$$\text{sim}(d, q) = \frac{P(R|d)}{P(\overline{R}|d)} \tag{4.4}$$

BM25 Model

BM25 implements document ranking using a bag-of-words retrieval function. It calculates the document similarity based on the query term existing in each document d and query Q. Here, the query Q is a set containing multiple keywords $[q_1, q_2, ...q_n]$. The BM25 model was introduced and improved in multiple research studies including [ZCT+04, LZ11]. The standard formulation of the BM25 term weighting function is:

$$\text{sim}(d, Q) = \sum_{i=1}^{N} \text{IDF}(q_i) \cdot \frac{\text{idf}(t, d, D) \cdot (k + 1)}{\text{idf}(t, d, D) + k \cdot (1 - b + b \cdot \frac{|d|}{\text{averdlen}})}, \tag{4.5}$$

where $averdlen$ is the average length of documents in the collection. k and b are two scalar parameters. The parameter k is a scalar parameter which calibrates the term frequency scaling. Setting the value of k to 0 reduces the BM25 model into a simple binary model in which no term frequency is considered, and increasing the value of k increases the impact of the term frequency. The parameter b is the other scalar parameter which controls the scaling of document length, whereas $b = 0$ means no requirement of document length normalisation. IDF(q_i) is the inverse document frequency of query term q_i, defined as [RJ88]:

$$\text{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}, \tag{4.6}$$

where N represents the number of documents in the dataset and $n(q_i)$ is the number of documents containing the specific query q_i. An open issue for the BM25 algorithm is how to determine suitable values of the parameters k and b. [MRS08] recommended the reasonable value of k for most tasks lies between 1.2 and 2 and b is set to be 0.75. [BMI12] suggests that the parameter settings, $k = 1.2$ and $b = 0.75$, are the default options for some existing industrial implementations. Based on these recommendations, we firstly apply the default parameter setting,

Table 4.2: Evaluated hyperlinking results using spoken information for ME13data (The results are presented as "RUN ID/MAP/tMAP")

	LIUM	LIMSI
TF-IDF	LIUM-TFIDF-13/0.1257/0.0768	LIMSI-TFIDF-13/0.1420/0.0814
BM25	LIUM-BM25-13/0.1285/0.0784	LIMSI-BM25-13/0.1501/0.0875

Table 4.3: Evaluated hyperlinking results using spoken information for ME14data (The results are presented as "RUN ID/MAP/tMAP")

	LIUM	LIMSI
TF-IDF	LIUM-TFIDF-14/0.0986/0.0496	LIMSI-TFIDF-14/0.1315/0.0712
BM25	LIUM-BM25-14/0.1067/0.0560	LIMSI-BM25-14/0.1331/0.0730

and investigate the target segment identification. In the last section of this chapter, additional research is reported in which we investigate the selection of the BM25 parameters to improve hyperlinking accuracy.

4.2.3 Experimental Investigation

In the remainder of thesis, we propose a set of experiments using various multimodal features. Our investigation aims to compare the difference between these features and corresponding parameter estimation. Thus, the motivation of this experiment is to select the optimal weighting model and ASR transcript to represent the textual feature in other experimental chapters, and we can focus on the discussion in multimodal feature analysis and avoid an overwhelming experiment data by involving both LIMSI and LIUM.

This experiment compares the hyperlinking results of the *TF-IDF* model and *BM25* model using ASR transcripts in terms of MAP. The ASR transcripts for ME13data and ME14data datasets are provided by LIMSI and LIUM, and were introduced in Section 3.4.1. The algorithms to identify target segments follow our previous researches presented in MediaEval 2013 and 2014 [CJO13, CJO14],

where a time-based sliding window was used to detect potential target segments. The window size is set to 120 seconds and the overlap is set to 40 seconds.

Table 4.2 and 4.3 show the hyperlinking performance using BM25 and TF-IDF algorithms. Both LIMSI and LIUM transcripts are investigated. We notice that all the runs using LIMSI transcripts in both collections (LIMSI-TFIDF-13, LIMSI-BM25-13, LIMSI-TFIDF-14, and LIMSI-BM25-14) achieve better MAP and tMAP values than the corresponding runs using LIUM transcripts. For example, in ME13data, LIMSI-BM25-13 has a MAP value of 0.1501, which is superior to LIUM-BM25-13 (0.1285). The MAP value of LIMSI-TFIDF-14 (0.1315) is also higher than the one of LIUM-TFIDF-14 (0.0986). Moreover, a similar conclusion can be achieved on the effectiveness of text retrieval models as BM25 works better than TF-IDF in terms of MAP and tMAP. For example, LIMSI-BM25-13 shows a better MAP value (0.1501) compared with its corresponding run LIMSI-TFIDF-13 (0.1420).

4.2.4 Discussion

The experiment results in Table 4.2 and 4.3 demonstrated the conclusion in [YH07] that BM25 works better than TF-IDF in terms of MAP. Moreover, the methods using LIMSI transcripts achieve better retrieval quality than those using LIUM transcripts. Thus, in the remainder of this thesis, our experimental design follows the conclusion achieved by the current experiments. It means that the hyperlinking retrieval model applying ASR transcripts involves LIMSI algorithm and BM25 retrieval model.

4.3 Identifying Target Segments

Section 3.2.2 explained that the target segments form the document collection for information retrieval in the hyperlinking framework, and that multimodal

feature analysis provides the description of the target segments. In Chapter 2, we introduced existing research from the MediaEval workshop and pointed out that researchers proposed various methodologies used to detect potential target segments from the video collections. However, this research on multimedia hyperlinking lacked detailed investigation of how the algorithms used to extract the target segments themselves influence the behaviour and effectiveness of the hyperlinking system. A hyperlinking process is usually a combination of target identification with other multimodal processing strategies. The retrieval results thus involve the combination of multiple components and do not show which target segment identification algorithm provides an optimal choice for the further investigation on multimodal analysis. In this thesis, identifying target segments is taken as an independent research topic. We introduce two state-of-the-art methodologies to segment a video for hyperlinking. Firstly, we describe the mechanism of these algorithms to segment the video stream using ASR transcripts. Next, experimental investigations show their hyperlinking performance associated with various parameters. Our motivation is to investigate the optimal settings for each algorithm, and determine which algorithm is most suitable for the multimodal processing to be used for experimental investigations in the remainder of this thesis.

4.3.1 Segment Identification using a Sliding Window

Using a sliding window is a simple and effective approach to identifying the target segments according to the review of MediaEval research in Chapter 2. A sliding window extracts fixed length segments from the video stream. The content of each extracted segment is determined by multimodal features contained within the time boundaries of the window. A sliding window has two parameters, the window size and overlap. The window size determines the length of a target

segment. The benchmark to assign the window size can be time-based or content-based. A time-based solution is relatively easy to implement compared with the content-based solution since there is no requirement to process lexical or other multimodal information in the former. The window size of the time-based solution ignores variations in the video content and sets a constant size for all the potential target segments in the collections. A content-based solution determines the boundary of segments using feature information. In Chapter 2, we reviewed a hyperlinking framework proposed in [GPKL14] which utilised machine learning algorithms to detect the end of a spoken sentence and indicated the window size according to the spoken sentence. Compared with the time-based solution, the primary difference of content-based is that the size of extracted segments varies since the window contents are related to lexical information when considering only ASR transcripts.

Another parameter associated with a sliding window is the overlap. The overlap means the common part between the two adjacent target segments and determines the distance between the start times of two adjacent segments. If the overlap is set to be 0, it means that the end time of one target segment is the start time of the next segment. Increasing the overlap means that two adjacent segments share more video content. In this case, if keeping the window size unchanged, the algorithm increases the number of extracted segments which can cover a variation of feature distribution in a video stream. The factors to determine the overlap between adjacent segments can be time-based or content-based as well.

The design of sliding window methods are based on the study of existing research proposed in the MediaEval workshop[2]. We categorise a sliding window as time-based or content-based. The details are introduced as follows.

[2]The review is presented in Chapter 2

Table 4.4: Implementation of time-based sliding window (pseudo code)

define the video V_i
define the fixed sliding window t_1 **and the overlap** t_2
define the start time $t_s = 0$
for the transcript T_i in V_i
    extract all text information located within t_s **and** $t_s + t_1$
    moving the sliding window $t_s = t_s + t_1 - t_2$
end for
save all extracted transcripts as potential target segments

Table 4.5: Implementation of content-based sliding window (pseudo code)

define the video V_i
define the number of text unit n_1 **and the overlap** n_2
define the start index n_s
define the transcript T_i in V_i
index all text units from 0 **to** $|T_i|$
for $n_s = 0$ **to** $|T_i|$
    extract all text unit indexed within n_s **and** $n_s + n_1$
    moving the sliding window $n_s = n_s + n_1 - n_2$
end for
save all extracted transcripts as potential target segments

- **Time-based Sliding Window**: This indicates a fixed window size and overlap in seconds. Therefore, the extracted segments share the same duration and the time interval between two adjacent segments is unchanged. The experimental design proposed in Section 4.2 utilises time-based sliding windows whose size and overlap are determined according to our previous investigation [CJO13, CJO14]. In this section, a grid search mechanism is applied to investigate whether there is an optimal design of time-based sliding window for ME13data and ME14data associated with ASR transcripts. Table 4.4 illustrates the pseudo code of time-based sliding window.

- **Content-based Sliding Window**: This method is motivated by [GP13] in which the authors employed a Decision Tree [GP14] to indicate the end of linked segment which demonstrated the importance of lexical information when segmenting ASR transcripts. Instead of employing a machine learning

algorithm, we directly utilise the segmentation information of ASR transcripts provided by LIMSI. The LIMSI algorithm splits the transcripts into sentences whose boundary is detected by the combination of audio features and semantic information [LG08]. Content-based sliding windows use each sentence as the starting point of a potential target segment, and this segment involves all spoken sentences located in a fixed duration. Table 4.5 illustrates the pseudo code for a content-based sliding window.

4.3.2 Experimental Investigation

In this section, we report our experimental investigation of strategies to identify target segments based on sliding windows as described in the previous section. Based on our preliminary experiments presented in Section 4.2, the segmentation algorithms are implemented using LIMSI transcripts. Indexing and searching on LIMSI transcripts is implemented by using BM25 algorithm, where default parameter settings are applied ($k = 1.20$ and $b = 0.75$), and the details of BM25 are explained in Section 4.2. Both the ME13data and ME14data datasets are used for this study. Hyperlinking retrieval is evaluated using the MAP and tMAP metrics presented in Section 3.4.3. Experimental results compare the hyperlinking performance of the time-based and content-based. Table 4.1 shows the abbreviations representing the methodologies used in the experimental runs. The parameter of both methods are introduced as follows:

- **Time-based Sliding Window**: Content analysis on a potential target segment extracts all the spoken words within the window. The size and overlap parameters determine the window size and overlap both in seconds respectively. The maximum window size is set to 150 seconds and the minimum size is set to 60 seconds. Various overlap values are examined in the ex-

Table 4.6: Results of the time-based sliding window solution (TSW-W) for ME13data. (The italic result is the best value achieved in Section 4.2.3, and the bold result is the best result in this table.)

	Size							
	60		90		120		150	
Overlap	MAP	tMAP	MAP	tMAP	MAP	tMAP	MAP	tMAP
10	0.1234	0.0887	0.1296	0.0863	0.1362	0.0821	0.1328	0.0705
20	0.1271	0.0865	0.1354	0.0865	0.1363	0.0792	0.1315	0.0626
30	0.1328	0.0892	0.1408	0.0890	0.1461	0.0899	0.1335	0.0668
40	0.1376	0.0901	0.1428	0.0901	*0.1501*	0.0875	0.1343	0.0730
50	0.1423	0.0914	0.1496	0.0918	0.1530	0.0871	0.1398	0.0747
60	-	-	0.1496	0.0910	0.1533	0.0889	0.1444	0.0787
70	-	-	0.1477	0.0926	0.1525	0.0845	0.1493	0.0745
80	-	-	0.1497	0.0942	0.1531	0.0893	0.1491	0.0742
90	-	-	-	-	**0.1537**	0.0909	0.1496	0.0754
100	-	-	-	-	0.1511	0.0951	0.1489	0.0779
110	-	-	-	-	0.1519	0.0948	0.1479	0.0846
120	-	-	-	-	-	-	0.1510	0.0829
130	-	-	-	-	-	-	0.1496	0.0860
140	-	-	-	-	-	-	0.1501	0.0872

periments. The minimum is set to be 10 seconds, and the maximum is determined according to the size of sliding window.

- **Content-based Sliding Window**: Define the start time of a sentence as T. The size of a sliding window continues to increase by checking the start time T' of the next sentence. If the time interval between T and T' is less than the pre-defined window size, the sliding window increases by merging the adjacent sentence. The process stops when merging an additional sentence would cause the segment size to exceed the maximum allowed size. In this experiment, we set the maximum window size to be 60, 90, 120, and 150 seconds.

Table 4.7: Results of the time-based sliding window solution (TSW-W) for ME14data. (The italic result is the best value achieved in Section 4.2.3, and the bold result is the best result in this table.)

	Size							
	60		90		120		150	
Overlap	MAP	tMAP	MAP	tMAP	MAP	tMAP	MAP	tMAP
10	0.1269	0.0713	0.1252	0.0779	0.1117	0.0652	0.1128	0.0663
20	0.1355	0.0749	0.1299	0.0829	0.1234	0.0685	0.1220	0.0655
30	0.1379	0.0816	0.1341	0.0810	0.1267	0.0734	0.1244	0.0716
40	0.1391	0.0849	0.1372	0.0809	*0.1337*	0.0730	0.1301	0.0756
50	0.1401	0.0843	0.1384	0.0833	0.1340	0.0764	0.1346	0.0724
60	-	-	0.1408	0.0841	0.1308	0.0749	0.1332	0.0769
70	-	-	0.1421	0.0860	0.1349	0.0727	0.1376	0.0782
80	-	-	**0.1445**	0.0872	0.1354	0.0712	0.1345	0.0775
90	-	-	-	-	0.1336	0.0790	0.1393	0.0726
100	-	-	-	-	0.1375	0.0779	0.1381	0.0792
110	-	-	-	-	0.1415	0.0776	0.1378	0.0781
120	-	-	-	-	-	-	0.1389	0.0766
130	-	-	-	-	-	-	0.1347	0.0742
140	-	-	-	-	-	-	0.1395	0.0711

Time-based Sliding Window

Tables 4.6 and 4.7 show experimental results using time-based sliding window in terms of MAP and tMAP. In the remainder of this discussion section, we use the ID in the format of "TSW-W-[Size]-[Overlap]" to represent the corresponding run showed in Tables 4.6 and 4.7 . The runs TSW-W-120-40 in both collections are defined in Section 4.2, and the corresponding MAP value is shown in italics in both tables. The greatest MAP values for both collections are shown in bold. In ME13data, the greatest MAP value is 0.1537, and the improvement is 2.4. In ME14data, the greatest MAP value is 0.1445, and the improvement over TSW-W-120-40 is 8.1%. The best tMAP values presented in both tables also increase. The improvement rate is 3.9% (from 0.0875 to 0.0909) in ME13data, and 18.8% in ME14data (from 0.0734 to 0.0872). Thus, we have an initial conclusion that

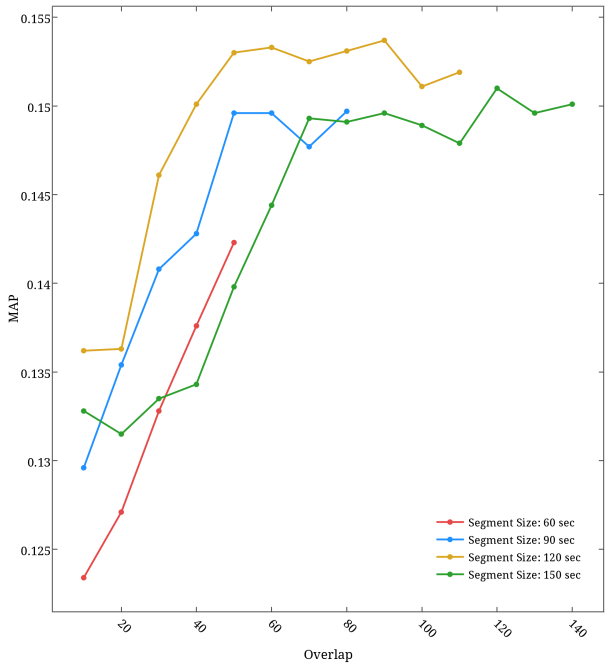


Figure 4.2: Investigation into the size of target segments for ME13data

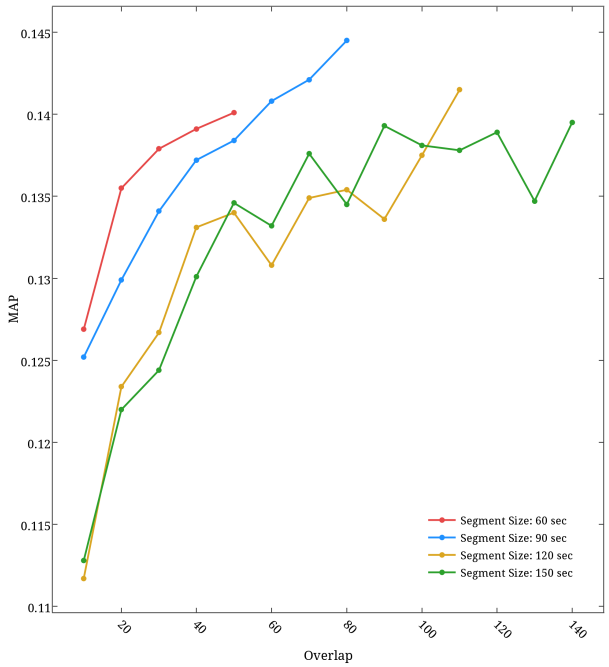


Figure 4.3: Investigation into the size of target segments for ME14data

using proper parameters for the time-based sliding window can improve the hyperlinking performance.

A research issue following the previous conclusion is whether we can determine the optimal parameters for the time-based sliding window solution. As two parameters, the window size and overlap, are involved, we analyse the relationship between them and the hyperlinking quality respectively. Firstly, we focus on how changing the overlap affects the MAP values.

Figures 4.2 and 4.3 illustrate the relationship between the overlap and MAP. Generally, when indicating the sliding window size, we can see that using a larger overlap achieves greater MAP values. In most cases, using a larger overlap can achieve an increasing MAP value compared to a smaller one. For example, when assigning a window size of 120 seconds, using a 110 second overlap in ME13data achieves MAP of 0.1519, which is greater than the one using 10 second overlap, 0.1362. In ME14data, the run using a 110 second overlap with a 120 second window size has MAP of 0.1415, which is also superior to the one using a 10 second overlap, 0.1117. We admit that in some cases, using a larger overlap decreases the hyperlinking quality. For example, the difference between TSW-W-150-110 and TSW-W-150-100 is -0.001 (from 0.1489 to 0.1479). However, the differences in MAP in these cases are statistically small. In conclusion, increasing the overlap can improve hyperlinking performance. When indicating the sliding window size, using a larger overlap can achieve better hyperlinking quality in terms of MAP and tMAP. The best retrieval in terms of the MAP value in ME13data is achieved with a 120 second window, while in ME14data, the greatest MAP and tMAP values are achieved when using a 90 second window. All the experiments demonstrate the poor effectiveness of using a small window size, since a 60 second sliding window always has a relative lower result. Besides, the experimental investigation illustrated in Figures 4.2 and 4.3 demonstrates the

Table 4.8: Evaluated results of the content-based sliding window solution (CSW-S) for ME13data and ME14data

RUN ID	MAP	tMAP	RUN ID	MAP	tMAP
CSW-S-60-13	0.1418	0.0963	CSW-S-60-14	0.1408	0.0716
CSW-S-90-13	**0.1609**	**0.1008**	CSW-S-90-14	**0.1505**	**0.0880**
CSW-S-120-13	0.1504	0.0977	CSW-S-120-14	0.1486	0.0798
CSW-S-150-13	0.1485	0.0961	CSW-S-150-14	0.1459	0.0722

difficulty of indicating an optimal sliding window size for both ME13data and ME14data.

By comparing the experimental results in Tables 4.6 and 4.7, we conclude that there are no definitive answers given to determine the optimal sliding window size using the time-based solution. The results in both tables agree that using a 60- or 150-second window cause decreasing hyperlinking performance, which means the size of sliding window should be moderate. However, the experimental investigation shows no evidence of the optimal size for the time-based solution.

Content-based Sliding Window

Table 4.8 presents experimental results using a content-based sliding window (CSW-S). We use the ID in the format of "CSW-S-[threshold]" to represent the corresponding run. The only parameter is the maximum size of the sliding window. The experimental results indicate that a 90-second threshold is better than the other two values in terms of MAP and tMAP for both collections. Furthermore, we conclude that CSW-S is superior to TSW-W. In ME13data, the best MAP of CSW-S is 0.1609, where the improvement is 7.2% over TSW-W-120-40-13. In ME14data, the improvement is 13.9% over the corresponding TSW-W-120-40-14. Experimental results also show the difference in the improvement between the optimal results for CSW-S and TSW-W. In ME13data, the optimal MAP value using CSW-S is 0.1609, which is higher than the one using TSW-W, 0.1537. In

ME14data, the greatest MAP value using CSW-S is 0.1505, which is superior to the one using TSW-W, 0.1445. Therefore, we conclude that when using appropriate parameters, the CSW-S strategy is superior to the TSW-W in terms of MAP and tMAP.

The aforementioned discussion shows that the TSW-W strategy failed to reach an agreement on identifying the optimal sliding window size, while the results shown in Table 4.8 indicate that the optimal window size for CSW-S is 90 seconds for both collections. Comparing the experimental results described in Tables 4.6, 4.7, 4.8, we offer a suggestion on the range of segment duration:

- The size of a target segment should be moderate as demonstrated in Tables 4.6, 4.7, and 4.8. When assigning an overshort (60 seconds) or overlong (150 seconds) size to a target segment, the experimental results in these tables show that the MAP value decreases compared with the best results achieved. The best results occur when the segment size is set to be 120 seconds in Table 4.7, and 90 seconds in Tables 4.6 and 4.8.

- From Table 4.6, the following effect of increasing a sliding window size can be observed. Firstly, the hyperlinking performance improves, and then decreases. We notice that the difference between CSW-S-60 and CSW-S-90 is higher than that between CSW-S-90 and CSW-S-120 in both collections. Moreover, the MAP value of CSW-S-150 is still higher than that for CSW-S-60, although these values are lower than the greatest one achieved by CSW-S-90. This means that the CSW-S strategy favours a larger window size to determine the size of potential target segments.

- There is an optimal threshold to determine the segment size for the CSW-S strategy, which is 90 seconds. It means that the CSW-S strategy is a better method for the target segments whose size is between 60 and 90 seconds.

Table 4.9: The number of created target segments created by various strategies

RUN ID	No.Seg	MAP	RUN ID	No.Seg	MAP
TSW-W-90-10-13	98,280	0.1296	TSW-W-90-10-14	154,724	0.1252
TSW-W-90-80-13	525,966	0.1497	TSW-W-90-80-14	841,602	0.1445
CSW-S-90-13	752,712	0.1609	CSW-S-90-14	1,222,244	0.1505

Therefore, we suggest that the optimal threshold for the CSW-S strategy should be around 90 seconds.

The previous discussion confirms the benefit of setting a larger overlap to improve hyperlinking performance. The TSW-W strategy determines the number of potential target segments by increasing the overlap between the adjacent segments, while the CSW-S strategy involves no parameter to the size of overlap between adjacent segments. In the other aspect, using each sentence as the header of potential segments, this solution initially creates a large amount of potential target segments. Table 4.9 shows the number of potential segments constructed by three representative strategies whose segment duration is around 90 seconds.

Table 4.9 shows the number of created segments. In the previous experiments, we illustrated that the best methodology in Table 4.9 is CSW-S-90 for both data collections, and the MAP value of run TSW-W-90-10 is the lowest. For the methodology TSW-W-90, using a 80-second overlap creates more segments than using a 10-second one, and we regard it as the primary reason for its better hyperlinking performance. The CSW-S strategy constructs a large number of potential segments compared with all other strategies. It utilises the lexical information, the boundary of spoken sentence, to identify the target segments. Its best performance in both data collections further demonstrates that both factors, analysing lexical information and creating sufficient target segments, are critical to multimedia hyperlinking.

4.3.3 Discussion

This section showed hyperlinking performance associated with two different approaches to extract target segments. Experimental investigation revealed that changing the strategy to determine target segment can influence hyperlinking results. We conclude that: 1) using a larger overlap benefit the improvement of hyperlinking performance by creating a large number of potentially interesting segments; 2) using the analysis of spoken sentences provided by LIMSI improves the hyperlinking performance; 3) the window size to determine potentially linked segments should be moderate. In the remainder of this thesis, according to these conclusions, we apply the content-based sliding window to extract target segments from multimedia collections.

4.4 Determining the Optimal Parameters for BM25 Algorithm

All the previous experiments use the BM25 algorithm to index and retrieve spoken information using the default parameters of $k = 1.20$ and $b = 0.75$. [BMI12] suggests this default setting and comments that further optimising parameter setting could improve the weighting quality of BM25. Thus, in this section, we investigate the impact of BM25 parameter assignment on hyperlinking retrieval. Our motivation is to select the optimal parameters k and b for spoken data retrieval on ME13data and ME14data and apply this conclusion to the experimental investigation in the following chapters.

4.4.1 Experimental Investigation

The experiments use the content-based sliding window solution to retrieve spoken information. Results for the default BM25 settings obtained in early experiments

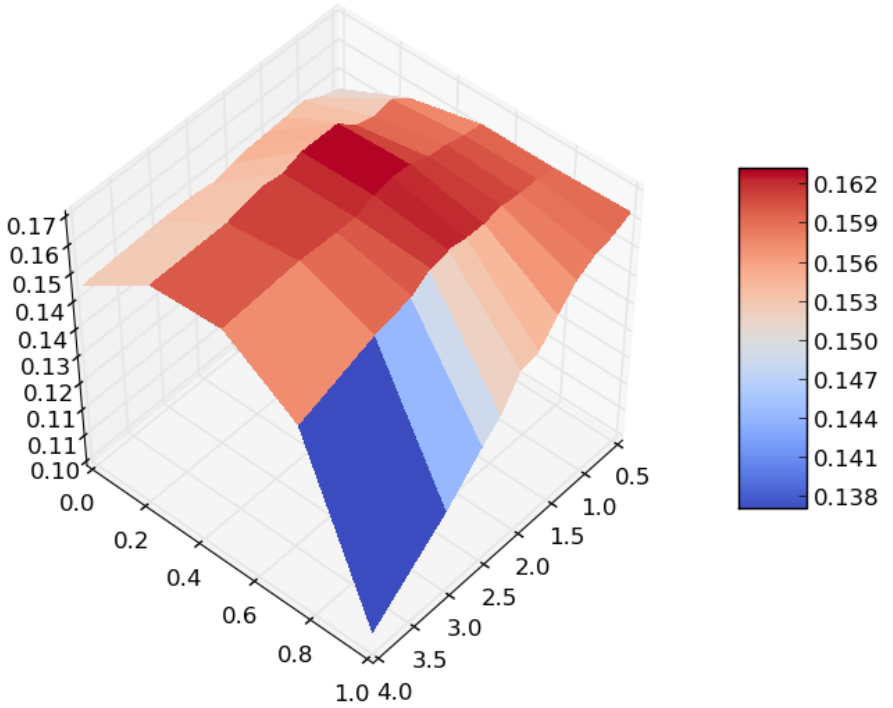


Figure 4.4: Hyperlinking performance using different BM25 parameters in ME13data

are shown in Table 4.8 (RUN ID: CSW-S-90-13 and CSW-S-90-14). k is typically found to be a float value which is greater than 0.0, and b is be defined as $0.0 \leq b \leq 1.0$. A grid search mechanism is applied to configure the optimal values of k and b. We choose a set of variables for k and b respectively. Hyperlink retrieval is designed according to the CSW-S strategy proposed in the previous section. BM25 parameters for the experiments are formed by assembling a set of possible combinations of k and b. The remainder of this section shows how the MAP values changes when utilising these combinations.

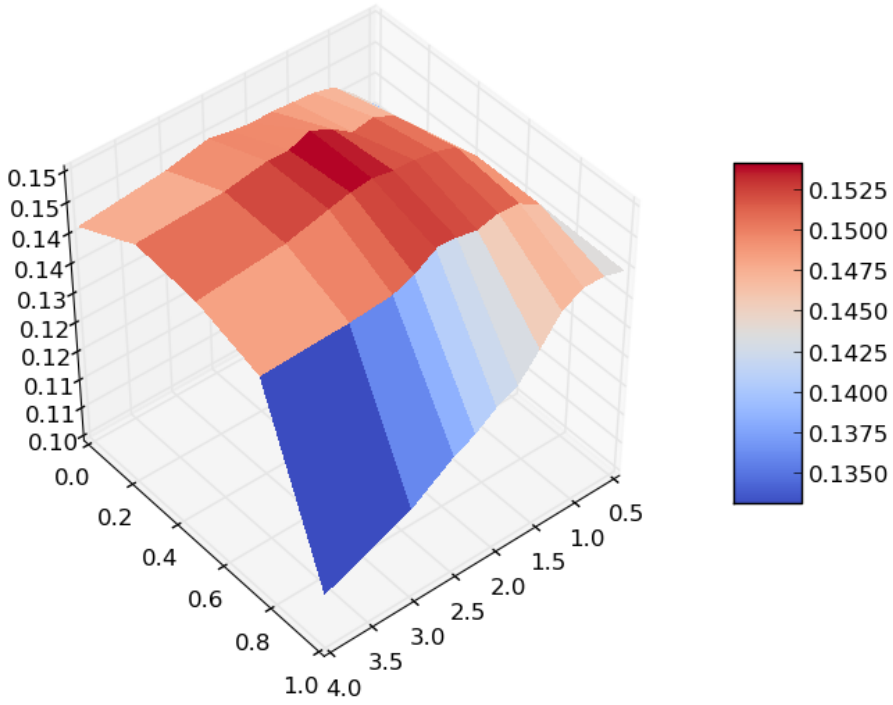


Figure 4.5: Hyperlinking performance using different BM25 parameters in ME14data

Figures 4.4 and 4.5 show hyperlinking retrieval using a content-based sliding window with various parameters for the BM25 algorithm. Each figure uses multiple colour levels to represent MAP values, from highest (red) to lowest (blue). In Figure 4.4, the optimal value of k is between 1.50 and 2.00 with $b = 0.5$. Setting $k = 1.75$ and $b = 0.5$ achieves the best MAP value of 0.1640. In Figure 4.5, the optimal value of k is between 1.50 and 2.25 and $b = 0.5$. The highest MAP value is 0.1546 when $k = 2.00$ with $b = 0.5$. The runs have agreement on an optimal value $b = 0.5$. However, there is some divergence on the optimal value of k. Based on the two sets of results, we suggest that k should be set between 1.75 and 2.00.

In Section 4.2, we demonstrated that using the default parameter assignment, *BM25* algorithm produced results superior to the *TF-IDF* algorithm for our hyper-

linking task. The current experimental results show that inappropriate parameter settings for BM25 algorithm can cause a decrease in results, with the values actually lower than those using the *TF-IDF* model. For example, when applying $k = 4.00$ and $b = 1.00$ for ME13data, the MAP value is 0.1052, which is lower than the worst MAP value (0.1257) presented in Table 4.2.

In conclusion, determining the optimal parameter for the BM25 algorithm is necessary for hyperlinking retrieval on spoken information. We found that, for BBC TV collections in hyperlinking, the optimal k can be arranged between 1.75 and 2.00 and the optimal b is around 0.5. In future experiments, considering that hyperlinking results using ASR transcripts can be seen as a baseline for multimodal feature analysis, we define $k = 2.00$ and $b = 0.5$ to implement the BM25 retrieval model for the ASR component.

4.5 Chapter Conclusion

In this chapter, we investigated the implementation of a hyperlinking framework using only text features. Our research concentrated on: 1) determining effective text-based indexing and searching methods to construct hyperlinks using spoken data; 2) investigating different strategies to identify potential target segments; and 3) indicating an optimal parameter setting for the BM25 retrieval model.

In Section 4.2, we compared two IR models to retrieve spoken information, *TF-IDF* and *BM25*. A hyperlinking framework was implemented by using the time-based sliding window with ASR transcripts (LIMSI and LIUM). Hyperlinking results show that using the *BM25* model achieves better retrieval results than using *TF-IDF* model. ASR transcripts produced by the LIMSI algorithm give a better description of spoken information. The experiments in Section 4.4 show that the BM25 model is superior to the *TF-IDF* when choosing appropriate parameters. Experimental results showed that a reasonable range of parameter settings for

BM25 is $k = [1.75, 2.00]$ and $b = 0.5$. Based on this finding, we determined to use these parameter settings for future experiments as $k = 2.00$ and $b = 0.5$. This conclusion is important because it allows us to concentrate on investigating other hyperlinking strategies with respect to using the BM25 algorithm on LIMSI transcripts in the remaining experiments.

Experiments investigated multiple strategies to extract potential target segments. Segment extraction methods were divided into two categories: using a time-based or content-based sliding window. To determine the appropriate parameters in different methods, a set of experiments was proposed. Evaluation results reveal that the window size and overlap between adjacent segments are both critical to improve hyperlinking retrieval. A large overlap has the advantage of covering sufficient multimedia information and decreases the chance of missing potentially relevant information. We conclude that the CSW-S strategy can create a huge number of potential target segments, and achieves a slight improvement in terms of MAP compared those using TSW-W.

The retrieval quality is not absolutely proportional to the increase of the window size. The TSW-W strategy failed to reach an agreement on the optimal sliding window size for both data collections. Experimental results in ME13data illustrate that a 120 second window achieved the greatest MAP value, while the conclusion changed to 90 seconds in ME14data. Therefore, we carried out a further investigation using other strategies. Experiments involving CSW-S whose duration limits were set to be 90 seconds achieved the best performance. We conclude that TSW-W can provide an untrustwrothy conclusion by segmenting the video stream without consideration of semantic information. Besides, the experiments show that a target segment should be moderate. We recommend 90 seconds as a reasonable threshold for the remainder of this thesis.

Also, the results indicate that the content-based sliding window improves the hyperlinking performance compared with time-based sliding window methods.

The former indicates potential target segments with a varying window size. In the experiment, content analysis is implemented in terms of extracting spoken sentences. Analysis of the experimental results shows that content-based analysis applied on spoken sentence (CSW-S) achieves the best performance for anchor-to-segment hyperlinking.

The contributions of the chapter are: 1) it showed how a segment-based hyperlinking framework is constructed. The experimental investigation used a complete hyperlinking process including target segment identification, segment indexing and searching. Different metrics to evaluate hyperlinking performance were presented. This process is also applied to future experiments in the thesis; 2) using text features, we propose how to identify the potential target segments. Experimental investigation shows that using lexical information can improve the quality of identifying target segments, and we apply this conclusion to the remaining experimental chapters.

The investigation described in this chapter only considers spoken information and ignores other multimodal features. When watching a video, the user's senses are guided by not only what they hear but also what they see. Thus, it is an essential research topic of how to efficiently involve multimodal features, especially visual features, to improve hyperlinking performance. Using multimodal features to provide a better hyperlinking service is the focus of the next chapter.

Chapter 5

Investigation on Multimodal Hyperlinking

5.1 Chapter Overview

The previous chapter focused on investigations of using ASR transcripts in our hyperlinking system. However, fully realising the value of the increasing number of multimedia archives available online requires users to engage in exploratory search behaviour to find content associated with multimodal features. To facilitate this, hyperlinks should be constructed based on the semantic information described by the text or visual contents of the archive. We expect that richer and more semantically meaningful hyperlinks formed using multimedia features can improve the user browsing experience by enabling enhanced navigation and recommendation. Thus, this chapter examines the creation of multimedia hyperlinks using spoken information in combination with other multimodal features. The primary goal of this study is to address the following research questions:

- RQ 3: How do other multimodal features except textual influence hyperlinking retrieval?

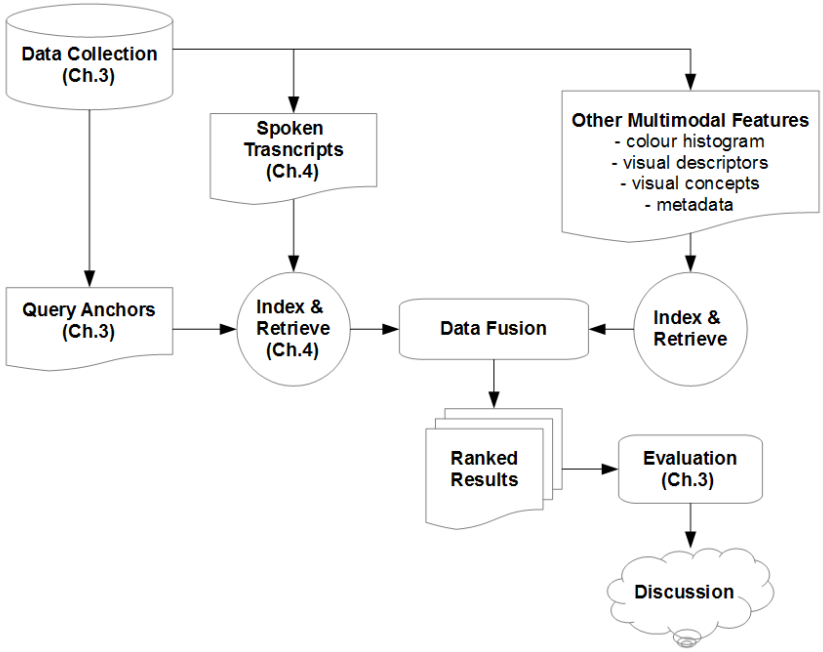


Figure 5.1: An overview of research design

- RQ 4: Can we improve data fusion strategies to integrate multimodal features for both ME13data and ME14data?

The chapter is structured as follows: Section 5.2 introduces our hyperlinking strategy using low-level visual descriptors; Section 5.3 describes our hierarchical hyperlinking model using high-level concepts; Sections 5.4 and 5.5 explore multimodal feature fusion in the creation of multimedia hyperlinks; and Section 5.6 concludes the chapter.

For readers' convenience, we propose Figure 5.1 and Table 5.1. Figure 5.1 illustrates a global view of experimental design proposed in this chapter, and Table 5.1 shows the acronyms of multimodal features and research methodologies. Reader can have a global view of our experimental design proposed in this chapter and a reference to check the acronyms in each experimental discussion.

Table 5.1: Acronyms in experimental investigation

Abbreviation	Description
ASR	LIMSI Transcript, content-based sliding window, BM25 weighting the window size is 90 seconds and the overlap is 30 seconds
CH	HSV colour histogram, correlation kernel
ORB	ORB descriptor, BoVW model, TF-IDF weighing
M	the video metadata as the video-level feature
C	the Oxford Concepts as the video-level feature

5.2 Hyperlinking using Visual Features

When watching video material, a user's interest is not only in what they can hear but also what they can see. This means that visual features are important to hyperlinking systems by representing what users see in this video. For this reason, it is essential to investigate the use of visual features in multimedia hyperlink construction. Chapter 2 introduced the classification of visual features as low-level and high-level. This section examines the use of two low-level visual features: HSV colour histogram and Oriented FAST and Rotated BRIEF (ORB) descriptor in hyperlink construction. We next present a detailed description of how to implement visual similarity analysis based on the video keyframes.

5.2.1 Low-level Visual Features

Colour Histograms

To explore the use of colour histogram in visual hyperlink construction for the BBC TV data collections (ME13data and ME14data), we apply colour histogram feature extraction using the HSV space introduced in Chapter 2 to process keyframes extracted from each shot in the video. Colour histograms have several advantages for image processing. The extraction algorithm is easily implemented and low-cost. When processing TV data, colour information can be used to recognise efficiently those images (keyframes) which share similar chromatic information.

A disadvantage of the colour histogram as a low-level feature is the lack of spatial information associated with simple colour bins. To address the weakness, we use a method that incorporates a spatial pyramid representation of video keyframes. Each keyframe is divided into 1×1, 2×2, and 4×4 grids. colour histogram analysis is applied to each grid to create a feature vector V. Each value v in this feature vector V is in the range $[0, 255]$.

There are multiple kernels to calculate the similarity between two colour histograms H_1 and H_2. To calculate the similarity score between two colour histograms, we use a kernel known as Correlation according to [CRM03, CEJO14], shown in Equation 5.1 and 5.2:

$$\text{score}(H_1, H_2) = \frac{\sum_{i=1}^{K}(H_1(i) - \overline{H}_1)(H_2(i) - \overline{H}_2)}{\sqrt{\sum_{i=1}^{K}(H_1(i) - \overline{H}_1)^2(H_2(i) - \overline{H}_2)^2}}, \tag{5.1}$$

$$\overline{H}_j = \frac{1}{K} \cdot \sum_{i=1}^{K} H_j(i), \tag{5.2}$$

where H_1 and H_2 represent two HSV histograms in vectors, and $H_j(i)$ means the ith descriptor in H_j. The length of the HSV vector is K which is determined by the space channels and the level spatial pyramid. In this thesis, the level of spatial pyramid is 3 (1×1, 2×2, and 4×4), the number of channels in the HSV space is 3, and the length of the HSV vector in each channel is 256. Thus, in this thesis, the value of K is 16,128 for each keyframe ($(1 + 2 \times 2 + 4 \times 4) \times 3 \times 256$).

Oriented FAST and Rotated BRIEF

The bag-of-visual words model (BoVW model) is widely applied to index and retrieve low-level visual descriptors in computer vision analysis [SZ03, SZ06, YJHN07]. Chapter 2 reviewed various low-level visual descriptors and the methodology to create a BoVW model from these descriptors. In the BoVW model, each image is represented by a sparse vector consisting of visual words

according to the occurrence of feature descriptors in the vocabulary. To create a BoVW model, we need to define: 1) what kind of low-level feature is used to represent the local features of keyframes; 2) how to build the vocabulary in terms of all the visual descriptors.

In Chapter 3, we explained that both ME13data and ME14data have a large number of video files (2,323 videos in ME13data and 3,520 in ME14data). It is obvious that a feature recognition algorithm will need to process a large number of potentially linked keyframes. Thus, a decision is required to balance the computing cost and the efficiency of feature annotation. Thus, we selected the Oriented FAST and Rotated BRIEF (ORB) descriptor introduced in Chapter 2 to describe the low-level features in video keyframes.

A strategy to create a visual vocabulary is to apply K-means algorithms to cluster K cluster points on existing visual descriptors. Each cluster centre represents a visual word. An open issue is how to determine efficiently the number of visual words (cluster centres) K. Chapter 2 reviewed some strategies to identify the optimal K experimentally. In this thesis, we apply the conclusion from [PCI+07], which suggested that K=20,000 is a reasonable value for the vocabulary size in multimedia information retrieval.

The procedures to index and retrieve ORB descriptors using the BOVW model are as follows.

- Use the ORB algorithm to detect visual descriptors in each keyframe. The algorithm is implemented using OpenCV ORB API[1] (Open CV Version 3.0.0). A 1,000$\times$32 matrix is generated to represent 1,000 ORB descriptors for each keyframe.

- Two vocabularies are created one for each of ME13data and ME14data respectively. In each collection, 1,000,000 ORB descriptors are randomly

[1]http://docs.opencv.org/trunk/doc/tutorials/features2d/akaze_tracking/akaze_tracking.html

picked for K-means clustering to concentrate K centre points (visual words in the vocabulary). K is assigned to 20,000.

- The ORB descriptors in each keyframe are matched to the corresponding vocabulary to detect the visual words. The FastANN algorithm [ML09] is used to accelerate the matching process.

- Removing stop words is an open issue for the BoVW model. Text-based retrieval has showed that removing the words with high frequency occurrence of in the data collection is critical to improving retrieval performance. [SZ03] applied a similar strategy to improve the quality of visual word indexing and searching. This paper suggests that visual words with high term frequency are less representative, and can be regarded as "visual stop words". In this thesis, we apply the same strategy proposed in [SZ03] to create the visual stop word list that involves all the visual words at top 5% of term frequency.

- Apache Lucene is used to index and search the visual terms. The TF-IDF model is used to calculate the similarity score.

5.2.2 Experimental Investigation

This section describes our experimental investigation using low-level visual features. The strategies include: 1) combining the hyperlink results retrieved by spoken transcripts with low-level visual features; 2) using the low-level visual features to rerank the top R results retrieved by the spoken transcript.

Combining Visual Features using Late Fusion

The experiment aims to create hyperlinks using the low-level visual features introduced in the previous section. The HSV colour histograms and ORB descriptors

using BoVW model are used to calculate the similarity scores. Experiments are conducted using both BBC TV collections, ME13data and ME14data. To identify target segments, we use the content-based sliding window solution introduced in Chapter 4 on LIMSI transcripts for both data collections.

The BM25 algorithm is used for content indexing and searching, with the parameter $k = 2.00$ and $b = 0.5$ following the conclusions presented in Section 4.3. Retrieved results using only spoken transcripts are defined as the baselines, named as **ASR-13** and **ASR-14** respectively.

The image at the middle time of a target segment or query anchor is selected as the corresponding keyframe. The following experiments are designed to investigate whether applying visual features can improve hyperlinking performance. Table 5.1 shows the hyperlinking strategies associated with the corresponding abbreviations. The **CH** and **ORB** results are retrieved according to Section 5.2. Runs using ME13data and ME14data are suffixed with the terms "13" and "14" respectively.

Data fusion integrates the hyperlinking results retrieved by the low-level visual descriptors with those retrieved by ASR transcripts. The following equation describes the CombSUM strategy to fuse multiple features according to Equation 2.1:

$$\text{Score}_{\text{fuse}}(q, s_j) = \sum w_i \cdot \text{Score}_i(q, s_j), \tag{5.3}$$

where w_i is the fusion weight for the ith multimedia feature, which can be spoken transcripts, colour histograms, or ORB descriptors. Each feature i achieves a normalised ranking score $\text{Score}_i(q, s_j)$ with respect to the query q and segment s_j. The ranking score is normalised according to the MinMax method outlined in Equation 2.5. In this section, all fusion weights are set to be the same weight

Table 5.2: Evaluating hyperlinking retrieval using low-level visual features for ME13data

RUN ID	Linked Feature	Fused Feature	MAP	tMAP
ASR-13	ASR	N/A	0.1631	0.0969
CH-13	CH	N/A	0.0507	0.0316
ORB-13	ORB	N/A	0.0652	0.0355
ASR-CH-13	ASR	CH	0.1377	0.0912
ASR-ORB-13	ASR	ORB	0.1462	0.0914

Table 5.3: Evaluating hyperlinking retrieval using low-level visual features for ME14data

RUN ID	Linked Feature	Fused Feature	MAP	tMAP
ASR-14	ASR	N/A	0.1546	0.0880
CH-14	CH	N/A	0.0314	0.0240
ORB-14	ORB	N/A	0.0428	0.0250
ASR-CH-14	ASR	CH	0.1318	0.0660
ASR-ORB-14	ASR	ORB	0.1282	0.0648

($w_i = 1$) for each feature, assuming an equal priority to determine multimedia content[2].

Tables 5.2 and 5.3 show a detailed comparison between the hyperlinking results using low-level visual features for ME13data and ME14data. From these results, it can be seen that hyperlinking results using only ASR transcripts perform better in terms of MAP and tMAP metrics, than those using low-level visual features (CH and ORB) in isolation or in combination with ASR transcript runs. This shows that when considering a single multimodal feature, spoken information has a higher effectiveness than low-level visual features when combined using a simple data fusion scheme.

In general, spoken information, colour histograms or ORB descriptors can be regarded as low-level features since they are directly extracted from multimedia resources. The latter two explain potentially relevant information in terms of computer vision rather than human recognition. [Yan06] pointed out that rather than the features explained in computer vision, users prefer to use those features

[2]The optimisation of fusion weight is introduced in the later sections of this chapter.

reflecting human cognition when searching or browsing multimedia resources. Colour histograms concentrate on detecting the background or objects sharing similar chromatic information, and ORB descriptors using a BoVW model detect the variance of low-level features. The experimental results illustrate that low-level visual descriptors can fail to describe cognitive information in a video shot. Users could regard two segments sharing the same news reporting room as relevant, meanwhile they both show a blue wall as the background. However, indicating the low-level feature "blue" is different from the cognitive information "a news reporting room with blue background".

As noticed above, fusing visual low-level feature results with ASR transcripts decreases the hyperlinking performance. According to [CEJO14], low-level features, although lacking cognitive recognition, can perform as a complementary to spoken information. This decrease performance could be caused by the use of the equal fusion weights. An assumption that multimodal features are equally important to contribute hyperlinking performance is apparently insufficient. Furthermore, our previous research [CEJO14] pointed out that applying a re-ranking strategy, even with an equal fusion weight, can significantly improve the hyperlinking performance, which is the central topic of the following section.

Re-ranking Strategy

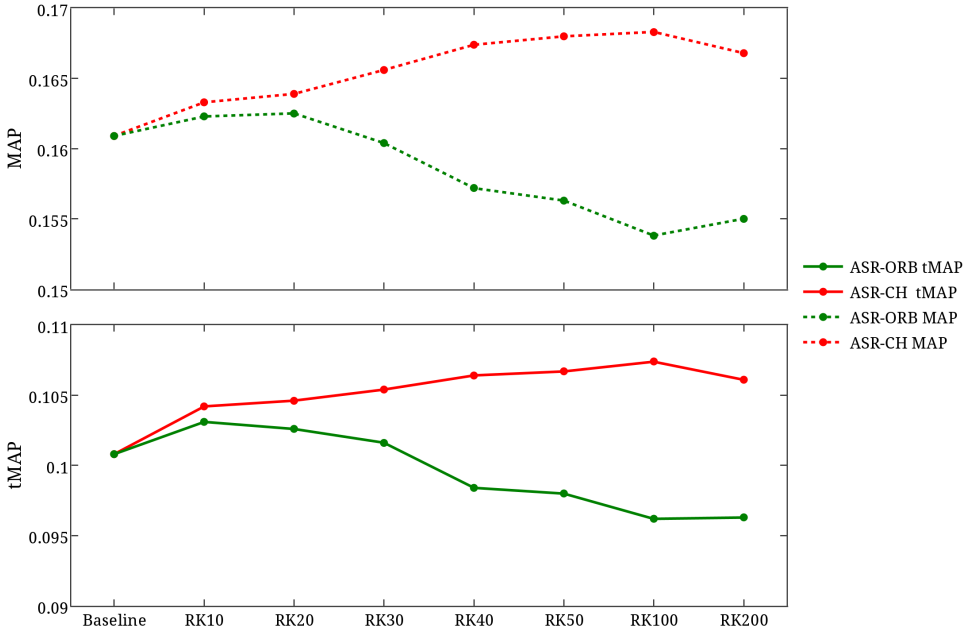
As reported in [CEJO14], a re-ranking strategy can be a simple and efficient methodology to improve multimodal-based hyperlinking. The strategy is: 1) create the initial retrieval list using spoken transcripts; 2) extract the corresponding low-level visual features; and 3) use late fusion to rerank the top R results by combining the similarity score achieved by low-level visual features and spoken transcripts. In this section, we investigate how this re-ranking algorithm performs in both ME13data and ME14data. The strategy uses low-level visual features to re-rank top R hyperlinking results retrieved by using ASR transcripts, where R is

set to be 10, 20, 30, 40, 50, 100, and 200. Fusing multimodalities is implemented using Equation 5.3, where the fusion weight w_i is set to be equal ($w_i = 1$).

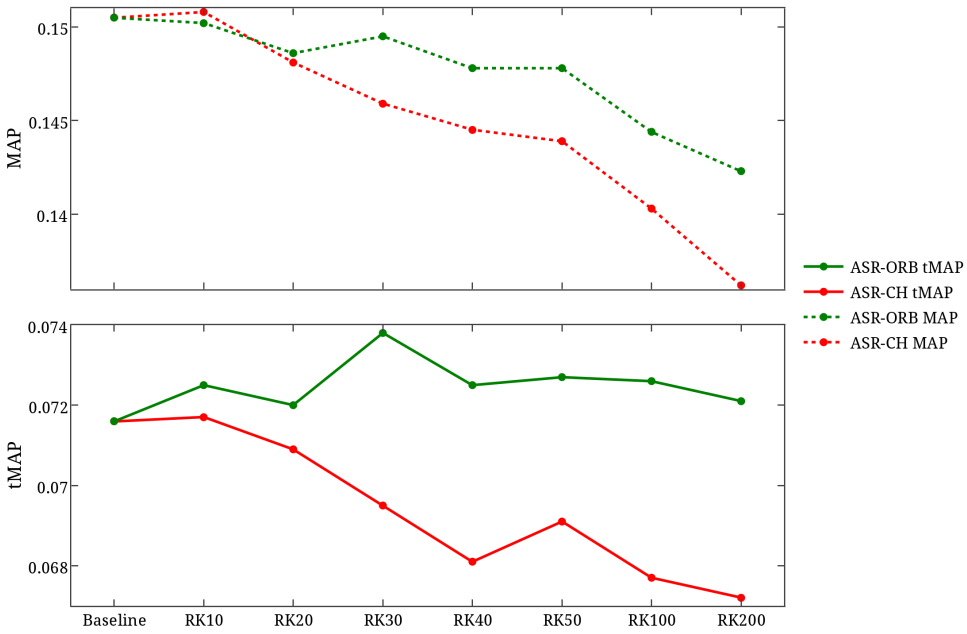
To illustrate the retrieval quality of the re-ranking algorithm, Figure 5.2 shows re-ranked results (**RK10** to **RK200**) associated with the baseline. The experimental results are inconclusive whether the re-ranking algorithm can improve hyperlinking performance. In the ME13data collection, the MAP and tMAP values increase at each re-ranking level and achieve the best result when re-ranking top 100 linked items. The best results achieve only 3.2% improvement over the baseline in terms of MAP (from 0.1631 to 0.1683). While for the ME14data, we observe that most results decrease compared with the corresponding baseline. The only increase of MAP for re-ranking of the top is from 0.1546 to 0.1548, but only 0.002 improvement can not demonstrate that the re-ranking strategy works in ME14data. In conclusion, it is clear from these results that this approach to re-ranking is an unstable strategy for improving hyperlinking quality.

Numerous factors influence a re-ranking method for hyperlink retrieval. One of them is the quality of initial retrieval using ASR transcripts. Figure 5.3 shows the results in terms of MAP value for **ASR-13** and **ASR-14**. It is clear that the retrieved MAP values for ME14data have a much larger variation than those for ME13data. For ME14data, 11 out of 30 queries achieve a very low MAP value, whose values are less than 0.05. The motivation for re-ranking the top R initial retrieved results is based on assigning relevant documents a higher rank based on other features. The effectiveness of a re-ranking algorithm, however, is obviously premised on the quality of initial retrieval. We speculate that the poor quality of retrieved results using spoken transcripts for ME14data is a critical factor in the unstable re-ranking of the results.

On the other hand, we should note that using equal weights for re-ranking fails to take into account the variance of multimodal features into account. In



(a)



(b)

Figure 5.2: Apply re-ranking algorithm (top 200) to fuse low-level features and ASR transcripts. (RK[R]: re-ranking top R results for (a) ME13data and (b) ME14data.

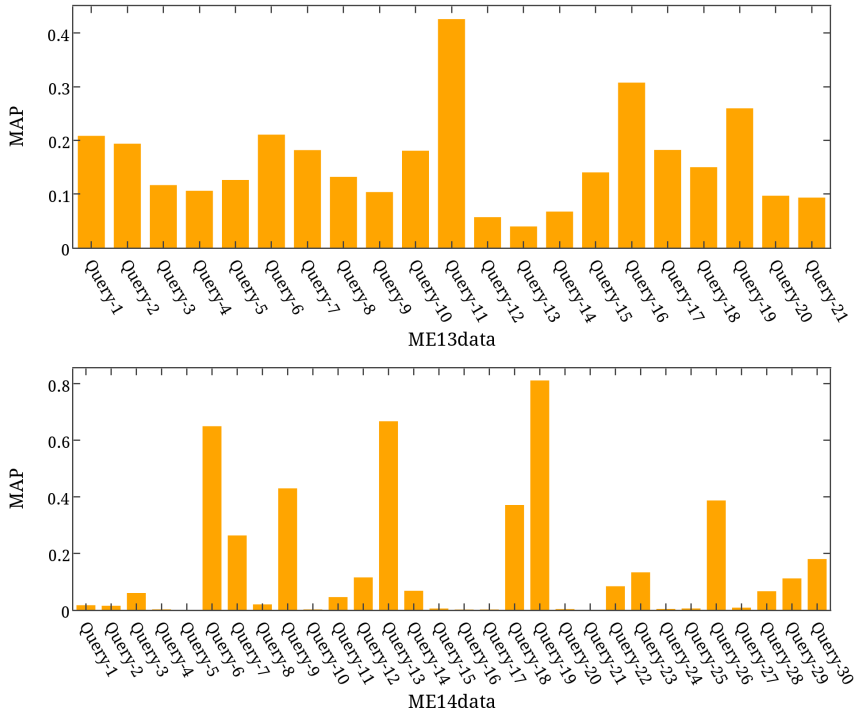


Figure 5.3: Hyperlinking performance for the initial retrieval for ME13data and ME14data

Table 5.2 and 5.3, the illustrated results show that spoken information is superior to low-level visual features in reflecting the cognitive sense of users in describing multimedia content. This means that spoken information should be dominant in multimodal fusion. Mathematically, the fusion weight for ASR transcripts should be larger than that for low-level visual features. Thus, to improve the re-ranking strategy for multimodal hyperlinking, we need to explore a strategy to assign an optimal weight for each multimodal field in the data fusion process.

5.2.3 Discussion

In this section, we have investigated the use of low-level visual features for hyperlink retrieval on the BBC TV data collections. Our methodologies included:

- Generate colour histograms to describe colour space information within the keyframes extracted from query anchors and target segments. Anchor-target similarity was determined using a correlation kernel.

- Extract ORB features to represent low-level visual descriptors shared by similar objects within keyframes. The BoVW model was applied to implement descriptor indexing and retrieval.

Experiments revealed that using only low-level features achieved relatively low hyperlinking quality, compared with the baseline using spoken information. The ineffectiveness of low-level features for hyperlink creation reveals that using multimodal features representative of cognitive information is important to search for potentially relevant links. The superiority of spoken information implies that when watching video shots, users prefer to understand the content from what they have heard. Thus, audio track information is critical for human annotation of the relevance of video shots. On the other hand, low-level visual features, lacking cognitive description, show a low effectiveness for retrieval relevant hyperlinks.

Next, we applied late fusion to integrate low-level visual features and spoken information. The previous experimental investigation showed worse results for both ME13data and ME14data. We believe that the primary reason for this is that we applied equal fusion weights to integrate multimodal features, which failed to reflect the diversity of the utility of the multimodal features for describing the relevant content.

Finally, we introduced a re-ranking strategy using late fusion. Results in Figure 5.2 show that this strategy improves hyperlink creation for ME13data, while for ME14data, the proposed solution did not achieve better performance. We believe that using equal fusion weights is again one of the reasons for the ineffectiveness of the re-ranking strategy for ME14data. A further experimental investigation also revealed that another factor was poor quality initial retrieval.

In conclusion, experimental investigation in this section suggests that the rest stage of our research on multimodal feature analysis for hyperlink creation should focus on:

- Use of multimodal features which are representative of cognitive information for a video shot.
- Optimisation of the combining weights for late fusion.
- Improving the initial results retrieved by spoken transcripts.

In the next section, we aim to use high-level features to improve hyperlinking performance. A further investigation is carried out on how to estimate linear fusion weights, and how to enrich the query content to increase the quality of initial retrieval for re-ranking.

5.3 Hierarchy Hyperlink Model

The previous section concluded that low-level visual features lacked representation of cognitive concepts. This section endeavours to improve hyperlinking retrieval using high-level features which describe cognitive information in multimedia resources. A hyperlink model, referred to as hierarchy hyperlinking, is proposed. The experimental investigation compares performance of our new hierarchy hyperlinking strategy with the previous investigations.

5.3.1 Segment-level and Video-level Features

Chapter 2 described the effectiveness of the summary information in multimedia retrieval. In this section, we use the summary information of the BBC TV collection to carry out the hyperlink retrieval. We assume that a target segment could be relevant to the video where this segment exists. In contrast with a video

segment, an entire video contains abundant information, including complete transcripts, metadata, or other semantic concepts. Therefore, we classify the multimedia features used in the hyperlinking system into two categories, referred to as segment-level and video-level.

All low-level features introduced in previous part of our investigation are segment-level, including the ASR transcripts, HSV colour histograms, and ORB descriptors. The identification of target segments determines what segment-level features are used and finally determines the hyperlinking quality. Video-level features, in contrast, are created by summarising the whole video stream, and are independent of the variance of target segments extracted from the video stream. Thus, all the target segments located in a video share the same video-level information. The following sections introduce two video-level features that can be extracted from the BBC TV collections.

Metadata

Our BBC TV collections provide manually annotated metadata for each video. The metadata consists of text descriptions of each video attributes, including its title, release date, coding information, and a brief description of the content. All the information is associated with the whole video, and therefore, can be considered as video-level features. In this thesis, we use the "description" information in the metadata to indicate a description of the actual video content. For our experiments, we use Apache Lucene to index and search metadata. Text analysis in metadata is applied according to the experimental hypothesis proposed in Section 3.4.4. The weighting model for indexing and searching is BM25, following the conclusion in Section 4.4.

High-level Concepts

The algorithm to create high-level concept is provided by the Vision Group at University of Oxford according to [CLVZ11]. A new high-level concept collection was created for the MediaEval Search and Hyperlinking task[3]. It contained a set of concept detector scores for 589 concepts in the video streams [CLVZ11]. The detectors were trained by downloading positive images from Google Images and learning their differences among negative images in the dataset using the libLinear toolkit [FCH+08]. We use the strategy proposed in [CEJO14, CJO13] to create a concept vector for each video. This strategy extracts all labeled keyframes within a video. Each concept value is determined by selecting the maximum value among those labeled keyframes. Thus, each video is represented by a concept vector. Cosine similarity is applied to calculate the similarity score between these concept vectors.

5.3.2 Use Hierarchy Hyperlinking Model

In the previous experimental investigation, the matching score between video segments (query anchors and target segments) was determined directly by the comparison of their segment content. The process was illustrated in Figure 5.4. To improve the hyperlinking quality, we propose an assumption: if a complete video is relevant to a query anchor, a target segment extracted from this video may be relevant to this query anchor. According to this assumption, a hierarchy hyperlink model divides the hyperlinking process into two steps. The first step is video-level searching. The video-level features are used to index and retrieve relevant complete videos whose scores represent the possibility that one or more related segments is contained within the video. The second step is segment-based hyperlinking. The hyperlinking framework seeks to identify potentially relevant

[3]The Search and Hyperlinking task uses the BBC TV collection.

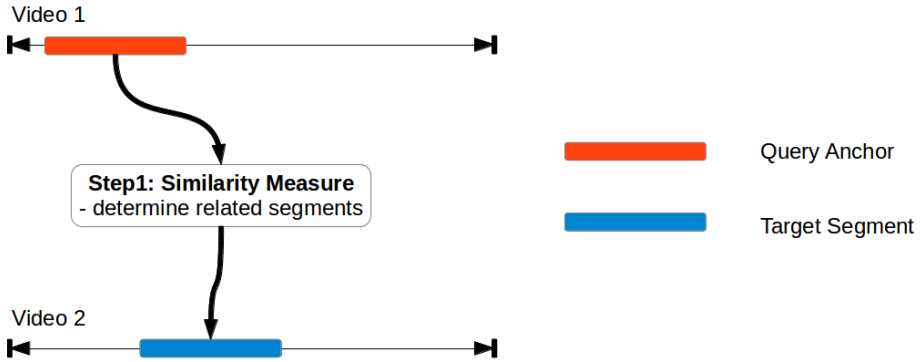


Figure 5.4: The segment-based hyperlink model

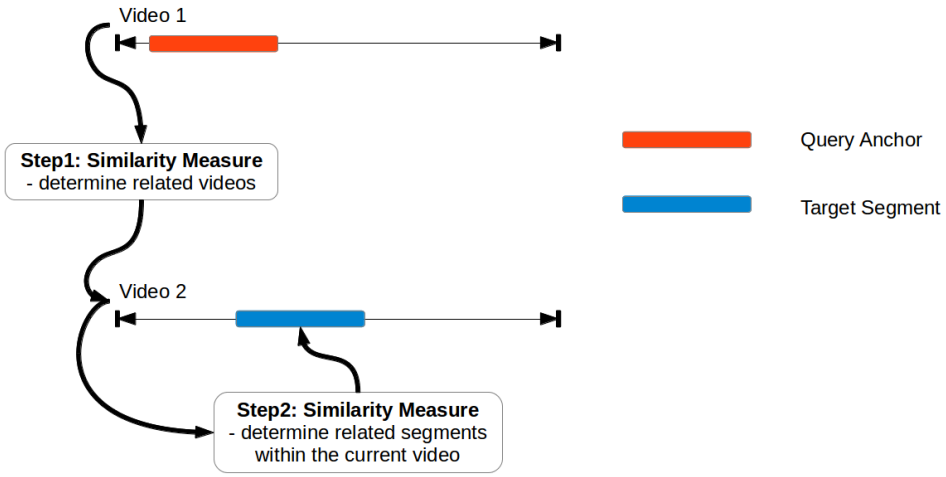


Figure 5.5: The hierarchy hyperlinking model

target segments within each video. The similarity score again is determined by the comparison of segment-level features extracted from the query anchor and the available target segments. Figure 5.5 illustrates this hierarchy hyperlinking strategy.

We propose to use a data fusion method to combine the video-level and segment-level hypotheses. Our combination method proceeds as follows. Define the target segment seg. The function Video(seg) indicates the video where a target segment seg is located. R_v represents the retrieval list of video-level search and R_s is the retrieval list of segment-based hyperlinking. A fused score $\text{Score}_{\text{hierarchy}}$ of the target segment seg is defined as:

$$\text{Score}_{\text{hierarchy}}(seg) = \text{Fusion Function}(R_v(\text{Video}(seg)), R_s(seg)), \tag{5.4}$$

where $R_v(\text{Video}(seg))$ returns the score of the video containing the current target segment seg in the video-level searching results, and $R_v(seg)$ returns the score of seg in the segment-level linking results. Using linear fusion, Equation 5.4 can be expressed as:

$$\text{Score}_{\text{hierarchy}}(seg) = w_v \cdot R_v(\text{Video}(seg)) + w_s \cdot R_s(seg), \tag{5.5}$$

where w_v and w_s are the fusion weights for video-level and segment-level ranked lists respectively. The ranking position of a linked segment is determined according to $\text{Score}_{\text{hierarchy}}$.

Equation 5.5 raises a fundamental research issue for the hierarchy hyperlink model: how to set suitable fusion weights. Multimodal features can have different contributions to effective multimedia hyperlinking. The hierarchy hyperlinking model needs to consider the diversity of video-level and segment-level features. This means that optimising the linear fusion weights is essential to our research. Our experimental investigation aims to address this issue in two steps. Firstly, in this section, our research focuses on whether the hierarchy model can improve hyperlinking performance with equal fusion weight. The experimental results are compared with those presented in the previous section. Our motivation is to investigate which features, segment-level or video-level, are more complementary to spoken transcripts when using a simple fusion strategy (equal weights). Secondly, in Sections 5.4 and 5.5, we investigate a strategy to determine the most suitable fusion weights to combine segment-level and video-level features.

Table 5.4: Hyperlinking retrieval using hierarchy hyperlinking

RUN ID	MAP	tMAP
ASR-13	0.1631	0.0969
ASR-M-13	0.2219/+36.05%	0.1181/+21.88%
ASR-C-13	0.1925/+18.03%	0.1081/+11.56%
ASR-14	0.1546	0.0897
ASR-M-14	0.2465/+59.44%	0.1159/+29.20%
ASR-C-14	0.2186/+41.40%	0.0934/+4.13%

5.3.3 Experimental Investigation

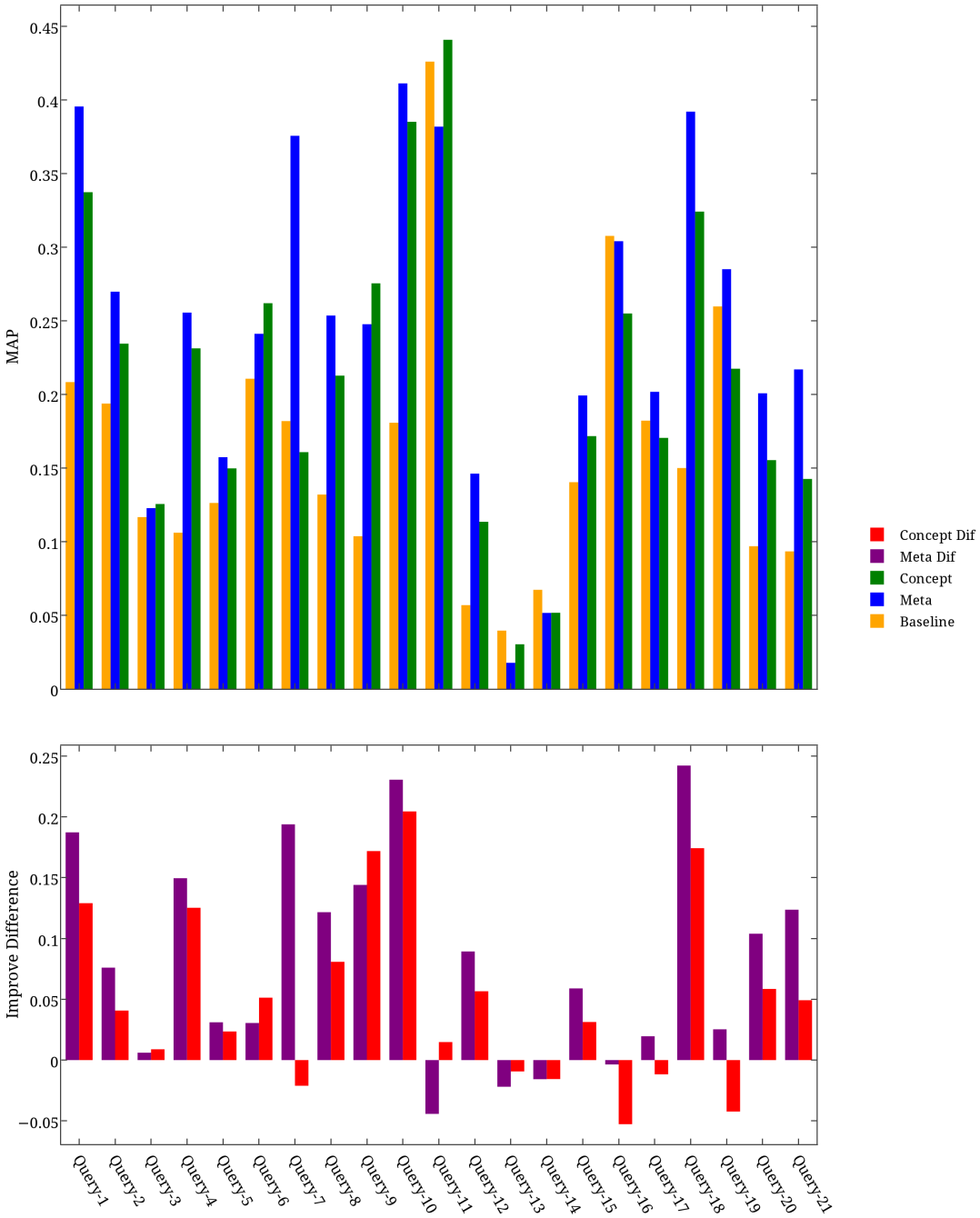
The experimental investigation in Section 5.2 showed that the spoken transcript is superior to other two segment-level features (colour histograms and ORB descriptors). Therefore, we use only spoken transcripts to perform segment-level hyperlinking. Table 5.1 lists the abbreviations used for the different experiments. We use Equation 5.5 to fuse video-level and segment-level features. Our motivation is to compare the hyperlinking results using the video-level features with the previous investigations. Thus, the fusion weights w_v and w_s are set to be equal ($w_v = 1$, $w_s = 1$). In the later sections, we will carry out a set of strategies to improve the estimation of fusion weights.

According to Table 5.4, the results using the hierarchy hyperlink mode (ASR-M and ASR-C) are better than the corresponding baselines (ASR) in terms of both evaluation metrics (MAP and tMAP). For the ME13data, the lowest improvement of MAP over the baseline is at least 18.03%, and the best is 59.44%. For the ME14data, the lowest improvement of MAP over the baseline is 59.44%, and the best is 41.40%. We can also observe improved tMAP values of ASR-M and ASR-C for both collections. Both video-level features can contribute better hyperlinking performance, even using equal weights on the linear fusion scheme. Thus, we conclude that the hierarchy hyperlink model using the video-level features can improve hyperlinking performance.

Table 5.4 shows that RUN **ASR-M-13** and RUN **ASR-M-14** achieve the greatest MAP values for the ME13data and ME14data. Recall that metadata information is manually created by the BBC, while Oxford visual concepts use spoken data and visual descriptors to build the concept dictionary automatically. It is obvious that human annotation can describe cognitive information of multimedia resources more accurately for a relevance determination. According to this experimental investigation, we conclude that hyperlinking retrieval using metadata outperforms those using visual concepts.

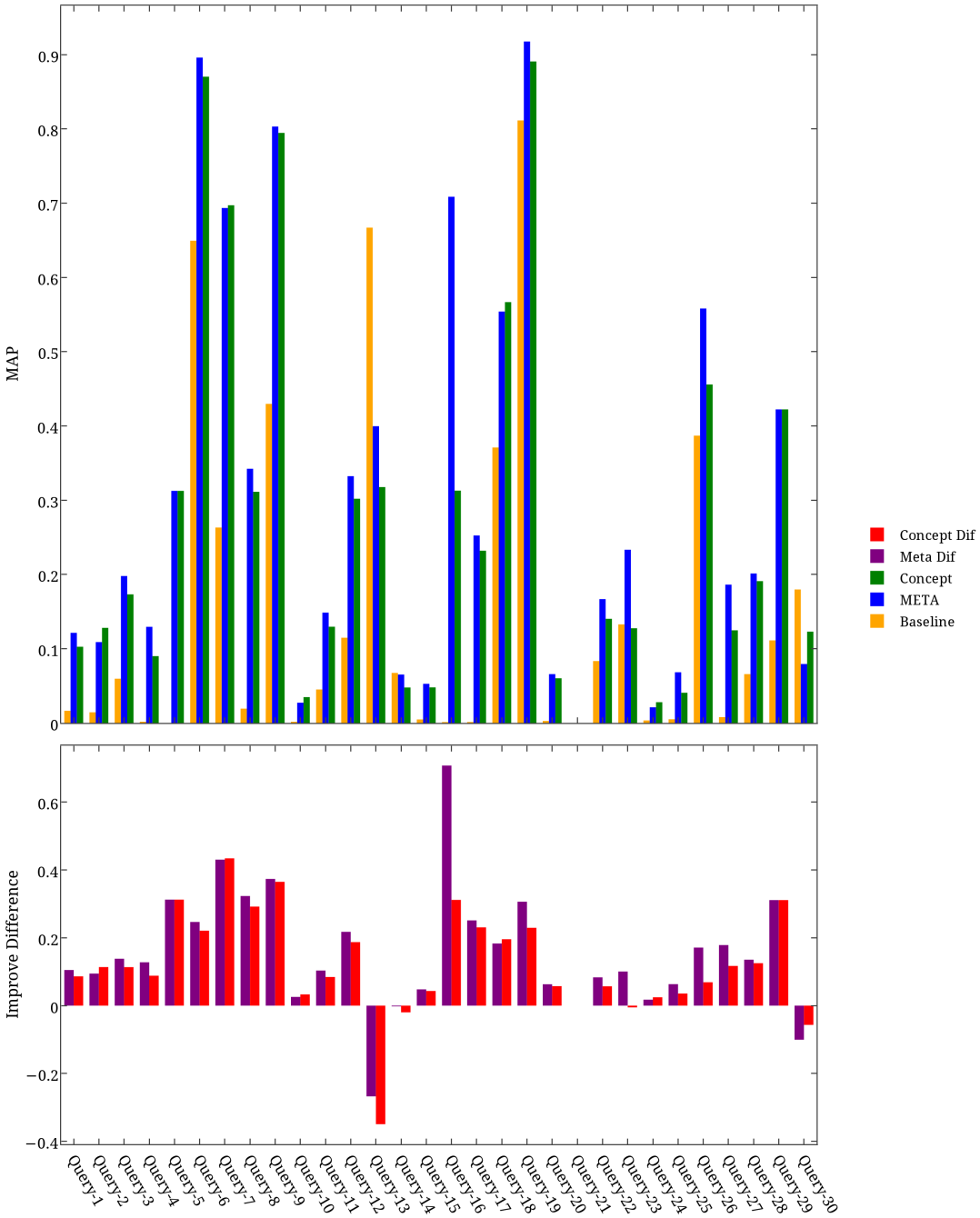
Figure 5.6 and 5.7 illustrate how the video-level features improve hyperlinking performance in terms of MAP on an individual query anchor. The figure shows the MAP values of all queries for both collections. The subfigure shows the improvement in each case between each query and its corresponding baseline. In Figure 5.6, 17 queries (21 queries in total) have increased the MAP values when fusing the metadata, and the greatest improvement is 0.2421 at Query-18. A total of 15 queries increase their tMAP values by fusing the visual concepts, and the best improvement is 0.2044 at Query-10. In Figure 5.7, 25 queries (30 queries in total) increase the MAP values by fusing the metadata, and the best improvement is 0.7075 at Query-16. A total of 25 queries increases the tMAP values by fusing the metadata with the best improvement of 0.4304 at Query-7.

It is obvious that the video-level features are more effective than the spoken transcripts to satisfy user requirements for hyperlinking. We take two examples, Query-13 and Query-16 in the ME14data, to describe the influence of video-level features on hyperlinking performance. The duration of the Query-16 is 83 seconds with 353 words in total. After removing the stop words, the query content contains 74 words including “India”, “China”, “camera”, “natural”, “train”, etc. This large number of words is still insufficient to represent the narrative of the video segments. The metadata describes the video as “Alesha Dixon looks at



(a)

Figure 5.6: Hierarchy hyperlinking performance on each query anchor for ME13data



(b)

Figure 5.7: Hierarchy hyperlinking performance on each query anchor for ME14data

the airbrushing of magazine photo". According to the crowdsourcing evaluation, most ground truth segments (21 out of 23) contain the keywords "Alesha Dixon" (person name) and "magazine photo" or similar concepts. Without the assistance of metadata, some linking results are shifted to other topics relevant to an introduction to China. It demonstrates that the video-level features have the advantage to provide essential keywords extracted from ASR transcripts. Thus, the results are improved by avoiding shifting to other irrelevant topics.

The video-level features improve most hyperlinking queries, however, with some exceptions. For Query-13, both solutions, using the metadata and visual concepts, decrease the performance of hyperlinking results. The duration of the query anchor is only 15 seconds, and the keywords include "aircraft", "carriers", "transforming Britain's ability to operate in hostile waters", etc. The metadata information is: "Start the day with the latest news, sport, business and weather from the BBC's Breakfast team". The video is the BBC Breakfast News, and the query segment talks about the reports of British military force. Using the metadata, most linked video shots point to target segments relevant to BBC breakfast news. These segments cover various topics that are quite different from "British military reports". Additionally, the visual concepts provide less assistance in improving hyperlinking performance. Spoken information containing specified keywords represents what users expect when watching this query video shot, and the hyperlinking results demonstrate this point.

5.3.4 Discussion

In this section, we introduced a novel hierarchy hyperlink model using video-level features. We investigate the use of two sources of video-level features: manual metadata and high-level visual concepts. A linear late fusion model, and equal fusion weights was used to combine matching scores from these sources

(video-level features). Following the positive results of our earlier experimental investigations, the experiments in this section used only ASR transcripts for matching at the segment-level.

Experimental investigation demonstrated that even using equal fusion weights, the hierarchy hyperlink model achieves better performance than our earlier experiments using only spoken transcripts for both ME13data and ME14data. We conclude that both features, the metadata and visual concepts, are effective at representing the potentially relevant information between those videos containing a query anchor and linked target segments. The experimental investigation also showed that when matching related videos, the metadata is superior to the visual concepts. We believe that the manually created nature of the information of the metadata by media professionals is the primary reason for this.

The experimental investigations in previous sections used a linear fusion model to combine segment-level and video-level features. As assuming that different features extracted from multimedia sources equally benefit hyperlink creation neglects the consideration of user preference when determining the relevance between a query anchor and a potentially linked segment. Thus, the remainder of this chapter will be focused on investigating the impact of varying linear fusion weights for combination of multimodal features or for the creation of relevant hyperlinks. Our methodologies use both supervised and unsupervised solutions to address the following questions: 1) can optimising the combining weights of late fusion improve hyperlinking retrieval; and 2) can the video-level features be superior to the segment-level features even after optimising the fusion weights?

5.4 Fusion Weight Estimation - A Supervised Solution

This section describes our investigation into multimodal fusion for hyperlinking performance using a supervised approach. The key issue examined here is the estimation of combination scores of the late fusion scheme is still an open issue. In the MediaEval hyperlinking task, significant training data for the BBC TV collection was not available until MediaEval 2014, when the workshop organisers released a training set based on the MediaEval 2013 experimental dataset. The availability of this training set encouraged further investigation of multimodal fusion analysis using supervised learning algorithms [CJO14].

Our research using a supervised approach concentrates on the use of machine learning algorithms to estimate late fusion weights for the different modalities. Our methodology is based on the theory presented in [MLD+14], where the authors pointed out that estimating fusion weights is equivalent to finding a linear axis that best separates the relevant and non-relevant documents in the ground truth dataset. The authors used Linear Discriminant Analysis (LDA) [YJL04] to estimate late fusion weights for multimodal features. Accoding to [MLD+14], we select LDA for the following reasons:

- LDA requires no estimation of parameters to build linear separation between relevant and irrelevant documents [MLD+14], which reduces the computation cost.
- [MLD+14] demonstrated the effectiveness of LDA in fusing textual and visual features in ImageCLEF collections[4], and our investigation has a similar target - fusing multimodal features in multimedia collections.

[4] http://www.imageclef.org/2009

In [CJO14], we applied LDA to determine the fusion weights between the metadata and ASR transcripts. The results confirmed that applying the LDA algorithm to fuse metadata and ASR transcripts achieved better results than the baseline, which used the spoken information (LIMSI transcripts) to link target segments. However, our research provided no evidence to show whether using supervised learning algorithms can effectively estimate multimodal fusion weights, compared with a relatively simple solution using equal fusion weights. In this section, we investigate whether using a supervised solution to estimate late fusion weights can improve linear fusion.

5.4.1 Linear Discriminant Analysis

Equation 2.1 shows the linear combination of multimodal features described in the previous sections. We assume that there are a total of R multimodal features to be combined. The combining score for the ranked list retrieved by the ith feature R_i is w_i. Our investigation focuses on how to determine appropriate coefficients w_i to improve linear fusion performance. A supervised solution, referred to as Linear Discriminant Analysis (LDA), is introduced in the following section.

LDA [Fis36] is an algorithm to find the linear combination of multimodal features for classification or dimensionality reduction. The effectiveness of LDA to determine linear fusion weights for the TRECVid multimedia collection has previously been demonstrated in [MLD$^+$14]. The core idea of LDA is to maximise the criterion of "between class variance" and "within class variance" to achieve the best linear separation for the ground truth [YJL04]. The coefficient vector achieving this separation is taken as the linear fusion weights for multimodalities.

We briefly introduce the procedure of LDA as follows. Define a multimedia hyperlinking ground truth as the training collection $\{T\}$. Define $\{X\}$ as the score vector of T, where $x_i \in \{X\}$ is the normalised score in the ranked list. $\{X_r\}$ is the

relevant vector of $\{T\}$ where r is 0 or 1, meaning that the video segment in the ground truth can be irrelevant or relevant to a specific query anchor. Thus, we have two scores vectors extracted from $\{X\}$ denoted as $\{X_0\}$ and $\{X_1\}$. $x_i \in \{X_0\}$ means that the ith element of $\{T\}$ is irrelevant to a specific query ($p(x_i|r=0)$) and its score is x_i. $x_j \in \{X_1\}$ means that the jth element of $\{T\}$ is relevant to a specific query ($p(x_j|r=1)$) and its score is x_j. The covariances of two score vectors $\{X_0\}$ and $\{X_1\}$ are Σ_0 and Σ_1, and the means are μ_0 and μ_1.

The concepts of "between class variance" S_b and "within class variance" S_w was proposed in [YJL04]. These are determined by the covariances and means of the score vectors, and the coefficient vector to combine multimodalities. LDA assumes that the coefficient vector for multimodal features is w where $|w|$ is the number of multimedia features involved hyperlinking, and both score vectors $\{X_0\}$ and $\{X_1\}$ follow the normal distribution. S_w and S_b can be determined using Equations 5.6 and 5.7:

$$S_b = (w \cdot \mu_0 - w \cdot \mu_1)^2, \tag{5.6}$$

$$S_w = (w^T \cdot \Sigma_0 \cdot w + w^T \cdot \Sigma_1 \cdot w). \tag{5.7}$$

The coefficients of linear separation can be calculated by maximising the criterion of between class variance and within class variance [YJL04]. Define the criterion $c = \frac{S_w}{S_b}$ as shown in Equation 5.8, based on Equations 5.6 and 5.7.

$$c = \frac{w \cdot (\mu_0 - \mu_1)^2}{w^T \cdot (\Sigma_0 + \Sigma_1)}. \tag{5.8}$$

A maximum separation can be accomplished by adjusting the weight vector as shown in Equation 5.9 [BG98]:

$$w \propto \frac{\mu_0 - \mu_1}{\Sigma_0 + \Sigma_1}. \tag{5.9}$$

The advantage of using the LDA algorithm to perform linear characterisation is that there no requirement for hyperparameter estimation. The coefficients w for the linear combination are only proportional to the ratio of between and within class variances. As the proportional change has no impact on the ranked results, in this thesis, we directly use the vector w output of LDA as the estimated linear fusion weights.

5.4.2 Experimental Investigation

This section describes our experimental investigation into combining segment-level and video-level multimedia features using a late fusion scheme. Table 5.1 lists the abbreviations used for our further experimental runs. A total of 13 strategies are applied to each data collection: **ASR_CH**, **ASR_ORB**, **ASR_CH_ORB** represent fusing ASR transcripts with other segment-level multimedia features, and **ASR_META**, **ASR_CPT**, **ASR_CPT_META** represent the hierarchy hyperlink solutions. The remainder of runs (**ASR_CH_META**, **ASR_CH_CPT**, **ASR_ORB_META**, **ASR_ORB_CPT**, **ASR_CH_ORB_META**, **ASR-_CH_ORB_CPT**, and **ASR_CH_ORB_META_CPT**) examine the strategies which combine other segment-level features to implement the hierarchy hyperlinking model.

The linear fusion scheme for multiple features is defined as Equation 2.1, where R_i indicates the list of ranked items retrieved by each multimedia feature, which can be *ASR*, *CH*, *ORB*, *META*, or *CPT*. The corresponding fusion weight for ith feature is w_i determined by the vector w output by LDA. The experimental investigation includes two parts. Firstly, we calculate the data fusion results using equal weights as the baseline. Secondly, we apply LDA to optimise the fusion weights.

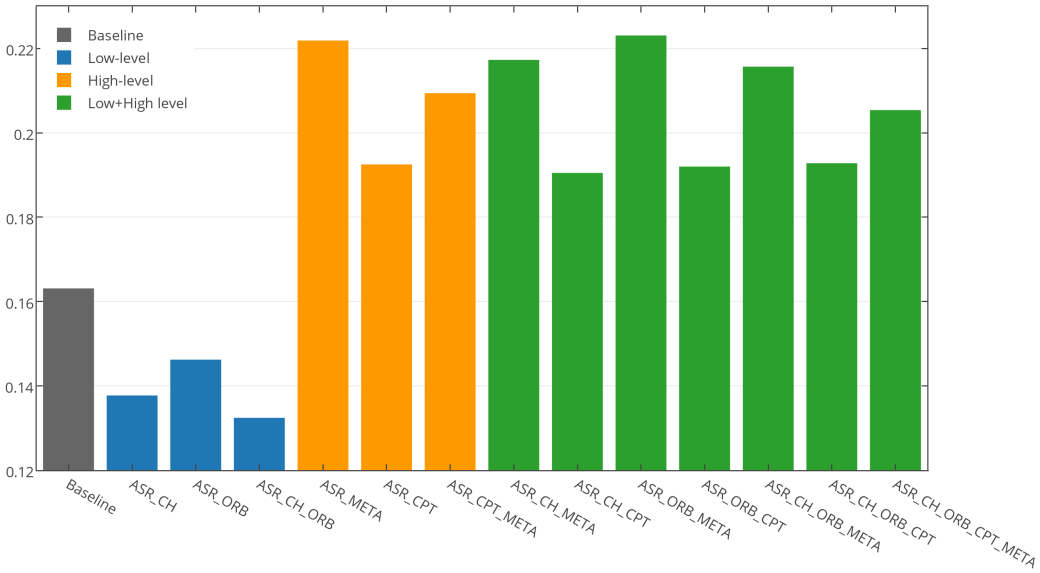
Multimodal Fusion using Equal Weights

Figure 5.8 shows MAP for multimodal hyperlinking results using equal fusion weights, where w_i is set to 1 for each feature. In total, 14 experiments are presented for both ME13data and ME14data, including one baseline (marked in grey) and 13 other runs as described in the previous discussion. We use the same baseline as in the experimental analysis in Sections 5.1 and 5.2. The blue bars represent fusing other segment-level features with spoken information. The yellow bars represent fusing the video-level features with spoken information, and the green bars indicate a combination of both video-level and segment-level features.

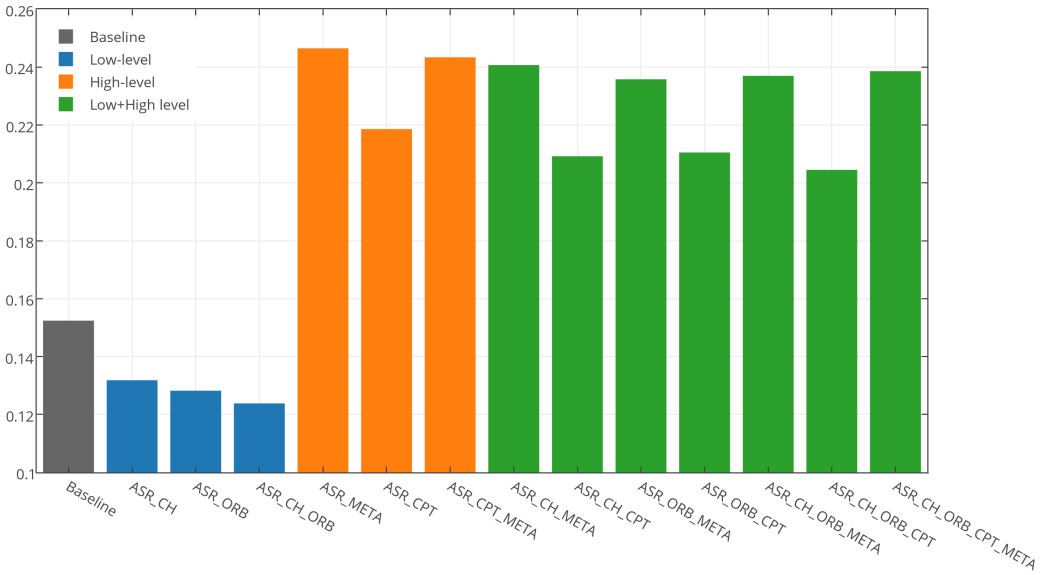
The experimental results demonstrate that directly fusing segment-level features can cause decreasing results. In both collections, fusing CH or ORB with ASR receives lower MAP values while fusing video-level features achieves an increase in performance. This observation was also seen in the experimental investigation presented in Section 5.2.2.

The hyperlinking results using video-level features (yellow and green bars) are superior to the baseline. However, the two data collections fail to reach an agreement on whether using segment-level features has a positive contribution to improving retrieval quality. In Figure 5.8 (b), ASR_META achieved the best MAP value 0.2465 compared with all other runs. In Figure 5.8 (a), ASR_ORB_META had the best performance, with a slight advantage, 0.2231, compared to ASR_META in second position, 0.2219. However, it is clear that for both collections, metadata information is the optimal feature to describe the video-level content. It is inconclusive whether combining both video-level and segment-level features can improve hyperlinking results.

The experimental investigation provides some basic conclusions for multimodal fusion on multimedia hyperlinking: 1) video-level features can increase the hyperlinking quality in all cases; 2) simply fusing segment-level features



(a)



(b)

Figure 5.8: Hyperlinking retrieval results fused by multimodal features using equal fusion weights in ME13data (a) and ME14data (b).

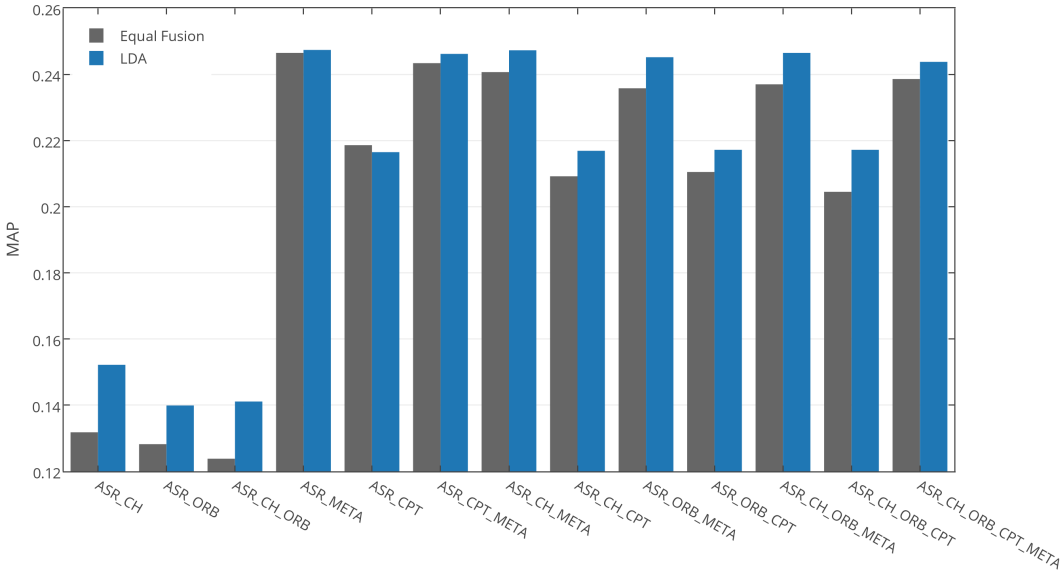
with ASR transcripts decreases hyperlinking performance; 3) the contribution of combining both video-level and segment-level features is inconclusive. These conclusions create the baselines for further experimental investigation and motivate our research to investigate: 1) whether using a supervised learning algorithm is superior to using equal fusion weights; and 2) whether the contribution of multimodal features can change after applying a supervised solution to estimate the fusion weights.

Multimodal Fusion using LDA

A challenge in the application of a supervised learning algorithm is that there is no reliable training data provided for the BBC TV collections. The MediaEval workshop proposed the Search and Hyperlinking Task in 2012 [EAL12] as a brave new task. After 2013 [EJC+13] and 2014 [EAO+14], the ME13data and ME14data were provided for testing and developing. However, no official training data was provided for ME13data, while ME13data itself served as the development set for ME14data. Therefore, the highest priority of our research is identifying a training data collection. In this section, we propose two solutions:

- Use the ME13data as the training set, and ME14data as the test set, which follows the official guideline provided by MediaEval workshop 2014 [EAO+14]. As a simple strategy, this method means there is no training set available for ME13data. All experiments are thus carried out only on the ME14data collection.

- We expect the experimental investigation to demonstrate the effectiveness of a supervised learning algorithm by applying LDA on both ME13data and ME14data collections. Therefore, cross-validation is used to create a training set for both data sets separately. Each collection is divided into three folds. Each fold contains the same number of queries, which means a total of 7



(a)

Figure 5.9: Hyperlinking retrieval results by fused multimodal features for ME14data. The fusion weights are estimated by using the LDA algorithm with ME13data used as the training set.)

queries located within each fold for the ME13data (21 queries in total), and 10 queries located within each fold for the ME14data (30 queries in total). Each collection is divided into three groups, and each group uses one fold as the testing set and two folds as the training set.

Figure 5.9 shows results obtained using LDA to estimate linear fusion weights for ME14data, when ME13data is used as the training set. Figure 5.10 shows the fusion results using cross-validation strategy within the corresponding data collection. All multimodal fusion results are marked in blue, and the baseline is marked in grey. In Figure 5.9, RUN ASR_CPT obtains lower MAP value than its baseline (from 0.2186 to 0.2165). Figure 5.10 illustrates some other cases with decreased results. In Figure 5.10 (a), the MAP value of ASR_CH decreases in Folders 1 and 2. In Figure 5.10 (b), the experiment results in Folder 3 show a lower

MAP value on RUN ASR_META. We conclude that using LDA fails to improve the hyperlinking performance for some multimodal features. This means neither of the proposed solutions can estimate proper fusion weights.

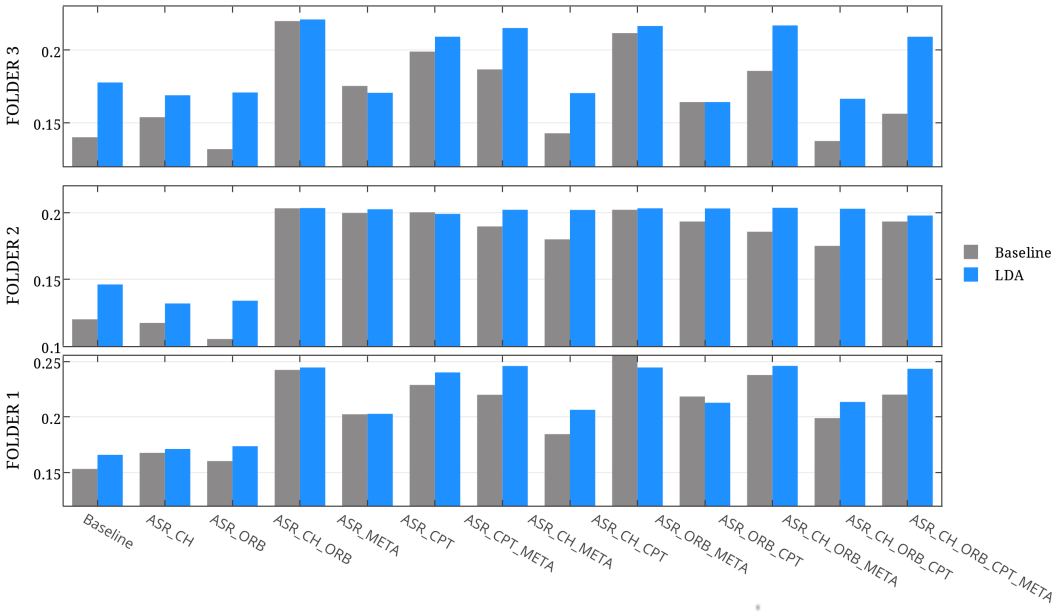
We analyse some potential reasons for the failure of using a supervised learning algorithm to estimate linear fusion weights. Firstly, LDA is probably not suitable to optimise the multimodal fusion weights for the BBC TV collection. Secondly, the optimal linear fusion weights could vary in their associations with each query anchor. Finally, the training data collection is inappropriate for optimising the fusion weights. The remainder of this section is dedicated to investigating the reason for the ineffectiveness of LDA.

In Figure 5.9, the runs ASR-META and ASR-CPT received conflicting results in terms of the multimodal fusion analysis. On one hand, using LDA improves the results of ASR-META. On the other hand, the MAP value of ASR-CPT is less than the baseline. Thus, in this section, we use the segment-level feature ASR and two video-level features META and CPT to recreate a binary linear fusion retrieval. According to Equation 2.4, we define this fusion process as follows:

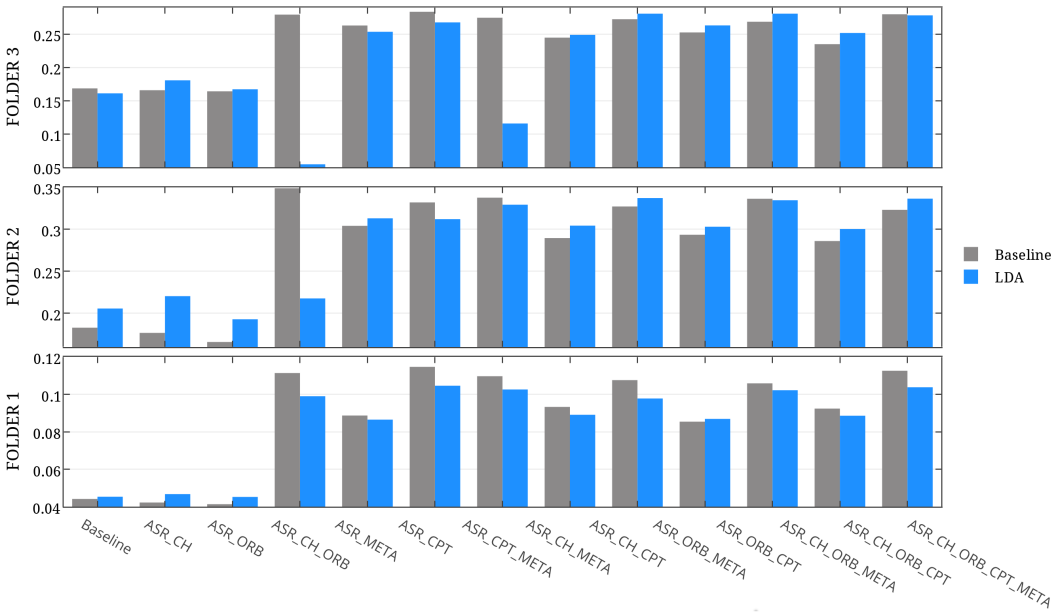
$$\text{Score}_{\text{fuse}} = w \cdot \text{Score}_{\text{ASR}} + (1 - w) \cdot \text{Score}_{\text{V}} \tag{5.10}$$

where w is the fusion weight for ASR transcripts. It is assumed to be normalised into the range $[0, 1]$. score_{V} is the ranking score retrieved by either META or CPT. Instead of using LDA, we manually assign w from 0.1 to 0.9, and evaluate the fused results in terms of MAP.

Figure 5.11 illustrates how the MAP value changes with various fusion weights for both ME13data and ME14data collections. The optimal fusion weights of META for the ME13data is around $w = 0.6$. Assigning w from 0.2 to 0.6 causes a slight variance in the MAP value. In general, it shows that spoken data in



(a)



(b)

Figure 5.10: Hyperlinking retrieval results for fusion of multimodal features for (a) ME13data and (b) ME14data using cross validation.

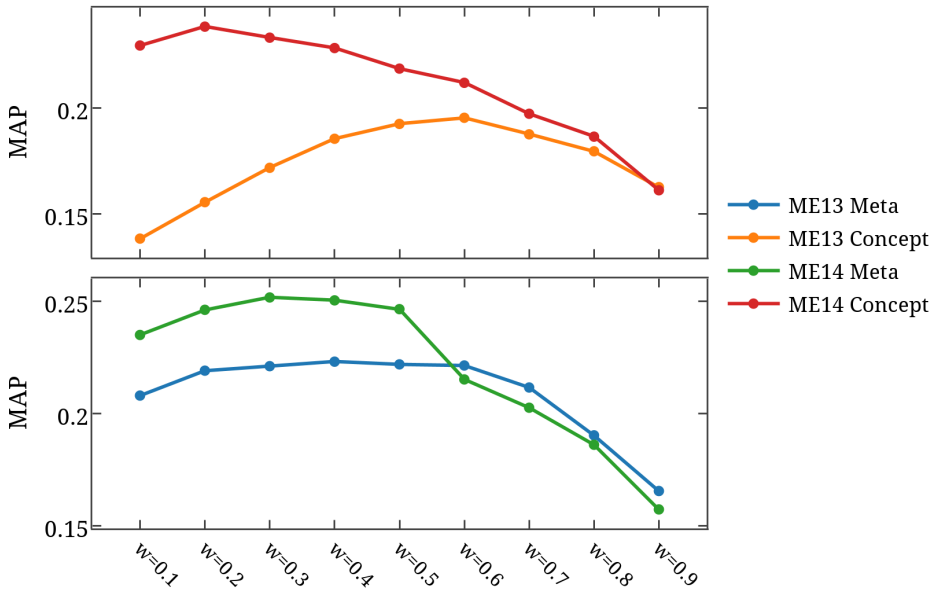


Figure 5.11: Investigation of the influence of fusion weights for the hierarchy hyperlinking model (w is the fusion weight of ASR).

the ME13data collection can provide sufficient information representing relevant information, and the binary fusion assigns a higher fusion weight to spoken transcripts. Meanwhile, for the ME14data, the figure indicates that assigning a lower weight to ASR provides a better fusion performance. This is in agreement with our previous assumption proposed in Section 5.3, which is that the video-level feature, especially for the ME14data, can be complementary to the query content extracted from spoken information. Hyperlinking results using the concept feature further demonstrates our conclusion. Figure 5.11 shows a clear disagreement on the optimal weights. The optimal weight for the ME13data is 0.2, and for ME14data, the value changes to 0.6.

The analysis explains why using a supervised learning algorithm decreases the effectiveness of hyperlinking multimodal fusion process when using ME13data as the training set and ME14data as the testing set. This reveals that the optimal fusion weights for multimodal features can vary in different data collections. Thus, the best linear separation for one data collection can not be suitable for another. This means that the ME13data collection is not suitable to be used as training data set for the ME14data, and vice versa.

5.4.3 Discussion

In this section, we proposed a supervised solution using LDA to estimate linear fusion weights for multimodal hyperlinking. We described two different experiments associated with two separate training collections. The results of these experiments revealed that a supervised solution does not provide a reliable estimation of multimodal fusion weights in the absence of appropriate training data. A further experiment demonstrated that the optimal fusion decision in the ME13data and ME14data could be different. The failure of cross-validation further shows that the primary issue to be addressed for optimising fusion weights is how to identify an appropriate training set.

The results raise the further consideration of multimodal feature distribution. An individual query could have a particular ratio between spoken information and visual features in describing a video story. Moreover, the hyperlinking ground truth used in the experimental investigation is constructed through crowdsourcing annotation. This means human cognition determines the multimodal feature distribution. We acknowledge that using fixed fusion weights can not represent the diversity of the contribution among various multimodal features in different query anchors. In conclusion, the supervised solution has two weaknesses. The first one is that there is no reliable training data provided for the BBC TV collec-

tions. The second one is that the optimal fusion weights change for each query, and using fixed weights calculated for a training collection does not represent the optimal contribution of multimodal features for each query.

5.5 Fusion Weight Estimation - An Unsupervised Solution

The previous section concluded that the absence of a suitable training set is a significant issue for estimating suitable linear fusion weights using a supervised learning approach. An alternative method for estimation of linear fusion weights is to use an unsupervised learning approach for an individual query anchor. Compared to a supervised solution, the potential advantages of unsupervised learning are:

- There is no requirement to select a training data set, which can overcome the negative effect of using an inappropriate training set.
- Users are expected to make different judgments on the importance of multimodal features when watching different video shots. Thus, assigning the fusion weights for each individual query could improve hyperlinking performance.
- The variety of multimedia features raises concerns about the use of a supervised learning algorithm. Additionally, a query anchor could be an arbitrary video segment in which users are potentially interested, and it is difficult to select a training data collection to represent all potentially interesting query anchors.

Therefore, this section describes our investigation into using unsupervised learning algorithms to estimate linear fusion weights. The algorithm used here

was originally presented in [Wil09], and is described in overview in the following subsection.

5.5.1 Maximum Deviation Method

We choose an algorithm, referred to as the Maximum Deviation Method (MDM) proposed in [Wil09], to optimise the fusion of multimodal features for hyperlinking retrieval . The reasons are:

- The previous section concluded that we can not indicate a reliable training collection for either ME13data or ME14data. MDM, as an unsupervised solution, requires no training data, which can address that issue.
- The author demonstrated its effectiveness in integrating multimodal features for the test collections provided by TRECVid 2003 to 2007 and Image CLEF 2007. Thus, we expect its effectiveness in our multimedia collections.

MDM is based on the assumption that a rapid change in the ranked scores of a retrieval list is an indicator of the potential importance of fusion [Wil09]. When fusing multiple retrieval lists, if we can observe a significant change in the scores in a retrieval list, this list potentially contributes more information for a data fusion process. On the other hand, if the change of the scores in a retrieval list is not significant, this list has less impact on the data fusion process compared with other retrieval lists. Thus, the "significant change" of the fused lists determines the fusion weights. To examine the "significant change" of the scores in a retrieval list, [Wil09] indicated the rank-based normalised scores as the benchmark. Given a retrieval list, MDM compares its normalised scores with its rank-based normalised scores to detect the "significant change" d as shown in Equation 5.11:

$$d = Max(\text{Score}_{\text{linear\_normal}}(seg) - \text{Score}_{\text{rank\_normal}}(seg)), seg \in R, \tag{5.11}$$

where d is the maximum difference between the normalised score sets of a retrieved results R. The function $\text{Score}_{\text{linear\_normal}}(seg)$ returns the MinMax normalised score of a video segment seg from R according to Equation 2.5. The function $\text{Score}_{\text{rank\_normal}}(seg)$ returns the ranked-based normalised score according to Equation 2.7. The fusion weight w_i for the ith feature is calculated as shown in Equation 5.12 [Wil09]:

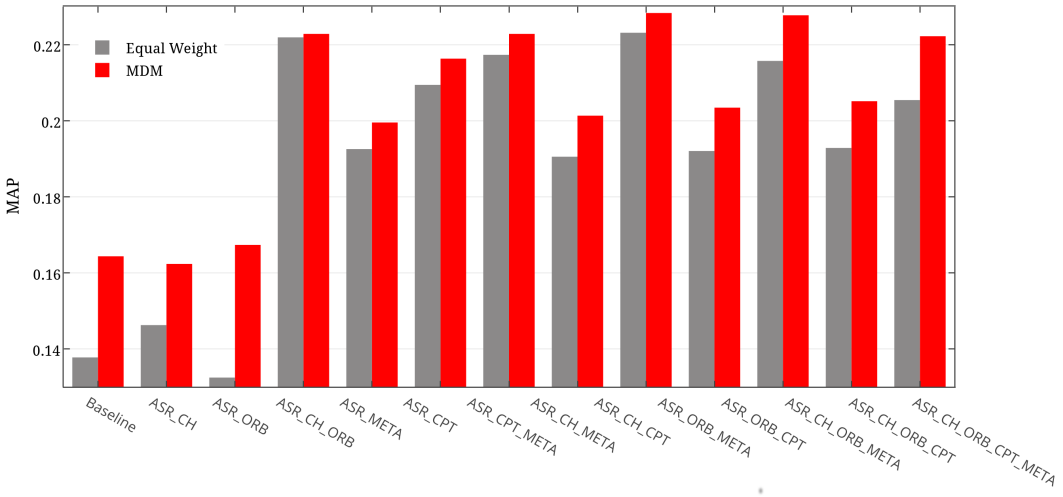
$$w_i = \frac{d_i}{\text{Rank}(d_i)} \cdot |R_i| \tag{5.12}$$

where $\text{Rank}(d_i)$ returns the rank position where the "significant change" is detected. $|R_i|$ is the size of the ith retrieval list.

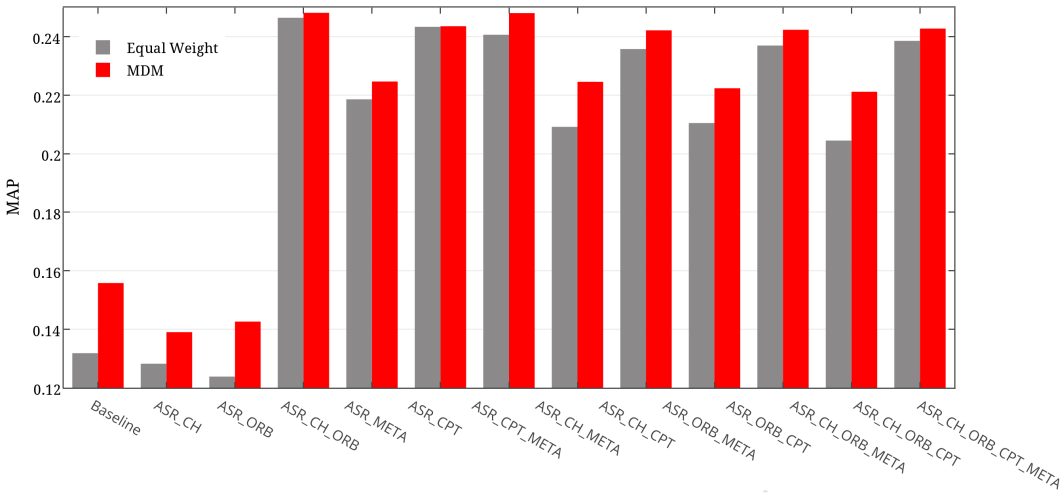
5.5.2 Experimental Investigation

This section describes our experimental investigation of the application of MDM to estimate linear fusion weights for multimedia hyperlinking. A total of five multimedia features (ASR, CH, ORB, META and CPT) are used to create 13 runs in our experiment. These are the same as those introduced in Section 5.4.2 and Table 5.1. The experiments compare MDM with two baselines: 1) using only ASR transcripts to link target segments, which was proposed in Sections 4.1 and 4.2; and 2) using equal fusion weights to combine multimodal features, which was proposed in Section 4.3. Both ME13data and ME14data collections are used in the evaluation.

Figure 5.12 shows hyperlinking results using MDM to optimise the linear fusion weights. The runs marked in red represent the strategy using MDM, and the runs marked in grey are the baselines where the fusion weights are equal for each feature. A clear conclusion is that using MDM can improve over linear fusion performance in both data collections. In Figure 5.12 (a), ASR_ORB_META achieves



(a)



(b)

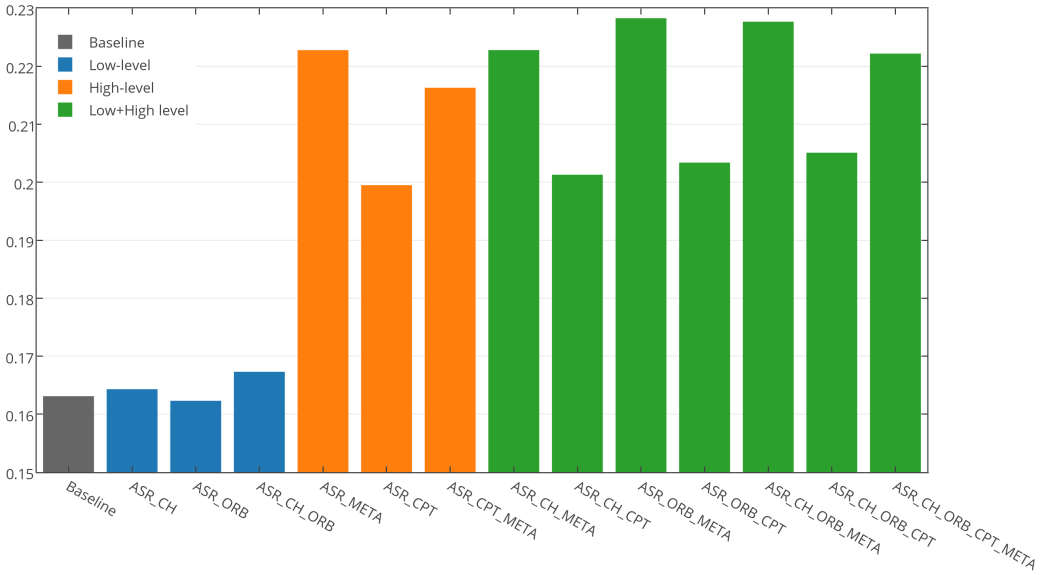
Figure 5.12: Hyperlinking retrieval results using the MDM algorithm to estimate fusion weights for (a) ME13data and (b) ME14data.

the greatest MAP value 0.2283. In Figure 5.12 (a), ASR_META achieves the greatest MAP value 0.2483. We observe that optimising fusion weights contributes more to the fusion between segment-level features (CH, ORB and ASR), while for the video-level features, the improvement is slight.

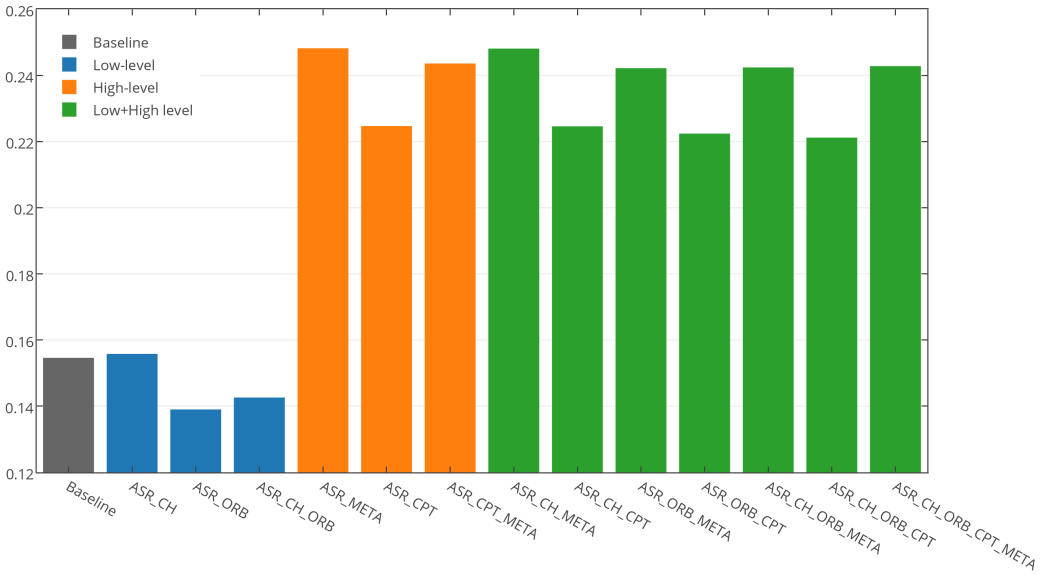
Figure 5.13 compares hyperlinking performance using MDM with the baselines using only ASR transcripts (marked in grey). We also use both segment-level (marked in blue) and video-level features (marked as yellow). The runs combining both are marked as green. Both figures demonstrate that video-level features are critical to hyperlinking performance. All the runs involving video-level features outperform the baselines in terms of MAP. In Figure 5.12 (a), ASR_ORB_META has the best MAP of 0.2283, and in Figure 5.12 (b), the best MAP is achieved by ASR_META, with a MAP of 0.2482, and a small improvement compared to ASR_CH_META (MAP: 0.2481). The fusion results demonstrate that the metadata outperforms the visual concepts. All the runs using META achieve better MAP values than those using CPT.

The segment-level features are less effective at improving the hyperlinking performance. We can observe that ASR_CH achieves better results in both data collections. The other runs (ASR_ORB and ASR_CH_ORB), however, can not improve the hyperlinking quality in both collections. The runs in Figure 5.13 (a) show the effectiveness of the use of multiple features. The run ASR_CH_ORB, which fuses CH and ORB features with appropriate weights, can improve hyperlinking retrieval compared with those using a single segment-level feature. In Figure 5.13 (b), we can observe an improvement from ASR_ORB to ASR_CH_ORB, but their results are still lower than that achieved for ASR_CH.

Figure 5.14 shows the linear fusion results using equal weights (marked in grey), LDA (marked in blue) and MDM (marked in red). We apply the same strategy used in Figure 5.10 to apply LDA and cross-validation to optimise the

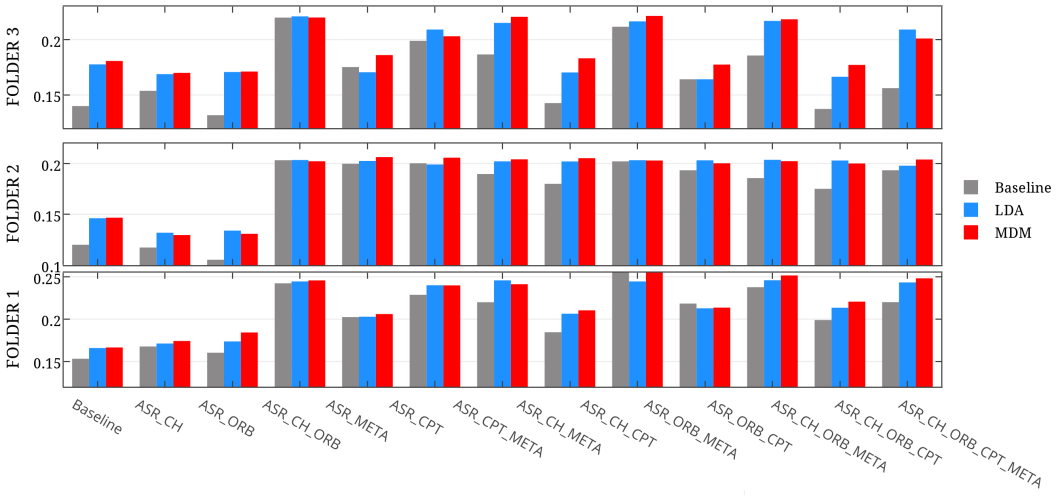


(a)

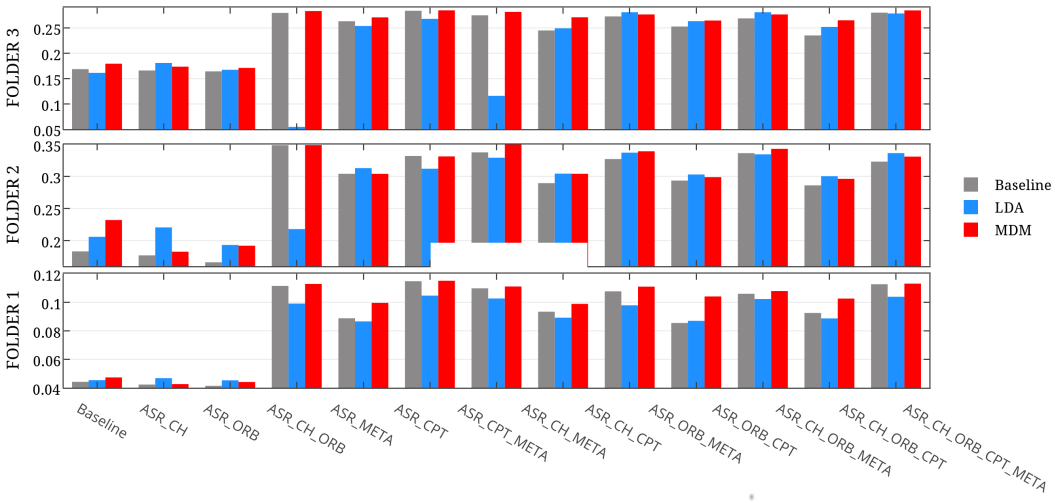


(b)

Figure 5.13: Investigation of multimodal fusion retrieval for (a) ME13data and (b) ME14data using the MDM algorithm to estimate fusion weights.



(a)



(b)

Figure 5.14: Comparison of hyperlinking retrieval results using LDA and MDM to estimate fusion weights.

fusion weight. This illustrates that a supervised learning algorithm cannot provide a reliable optimisation of fusion weights for this task with the available training data, while using MDM can achieve an improvement in terms of MAP. In some cases, however, we can observe that using LDA can achieve the best MAP. For example, in Figure 5.14 (a) FOLDER 3, the MAP of ASR_META using LDA is 0.2211, which is better than those of two other solutions (baseline: 0.2200, MDM: 0.2201). In Figure 5.14 (b) FOLDER 2, ASR_ORB using LDA also achieves better MAP (0.2204) compared with ASR_ORB using equal weights (0.1767) and ASR_ORB using MDM (0.1824). This suggests that if we could identify a proper training data set, a supervised solution might be more effective at improving hyperlinking performance.

5.5.3 Discussion

In this section, we investigated using the MDM algorithm, an unsupervised learning algorithm, to estimate linear fusion weights for multimodal hyperlinking. The advantage of this solution is the ability to identify a particular fusing weight for each query. Experimental investigation showed that using MDM can provide a reliable optimisation for multimodal fusion. Thus, in the remainder of the thesis, we continue to use MDM to estimate fusion weights for multimodal features.

5.6 Chapter Conclusion

This chapter described our investigation into using multimodal features to create hyperlinks across the BBC TV collections. Multimedia features include:

- HSV colour histograms, a low-level visual feature, extracted from video frames with multiple grid levels,

- the ORB descriptors, a low-level visual feature, extracted from video frames and used to create bag-of-visual-words,
- the metadata, a high-level textual feature to describe video content, provided by the BBC,
- the visual concepts, a high-level visual feature to describe the semantic information in video frames, supplied by University of Oxford,

In Section 5.3, we categorised these features as segment-level or video-level. A segment-level feature dynamically changes in association with the identification of target segments, while the video content determines the video-level features so that all the segments extracted from the corresponding video share the same video-level features.

We proposed a set of methodologies using multimodal features to create hyperlinks, and these features are combined using late fusion. For segment-level features, we sought to improve the hyperlinking quality by applying a re-ranking strategy. For the video-level features, we establish a hierarchical hyperlinking model to integrate the video-level and segment-level features.

Our experimental investigation revealed that the visual segment-level features, including colour histogram and ORB descriptors, are insufficient to represent cognitive information in the video stream when compared with spoken information. However, when we used the re-ranking strategy that was effective in our previous investigation [CEJO14] on the blip.tv data collection, this solution caused unreliable hyperlinking for ME14data collection. We conducted a further comparison between the quality of the initial retrieval using spoken information for both ME13data and ME14data collections. This revealed that the poor initial retrieval was one of the primary issues in the decrease in re-ranking performance. Furthermore, another potential reason for decreasing results is the use of equal fusion weights to combine different segment-level features.

Section 5.3 described a strategy to improve hyperlinking performance referred to as the hierarchy hyperlinking model. Our motivation for this model was that applying video-level features can be complementary to content analysis of potentially linked target segments. This model uses both segment-level and video-level features. Mathematically, multimodal combination can be implemented using a linear late fusion model. The linear fusion model fuses the initial retrieval using the metadata, visual concepts, and spoken information. Our experimental investigation showed that the hierarchy hyperlink model can improve hyperlinking performance in terms of MAP and tMAP, even when using equal fusion weights.

The research in Section 5.4 and 5.5 focused on optimising linear fusion weights to combine hyperlinking retrieval from multimodalities. In Section 5.4, we applied a supervised solution to determine the optimal separation of the training data collection using Linear Discriminant Analysis (LDA). The experimental investigation, showed that this supervised solution failed to improve the hyperlinking effectiveness. A further investigation indicated the difference of the optimal fusion weights between the ME13data and ME14data collection, which meant that there was difficulty in identifying a well-designed training set for both collections. Additionally, from the experimental results in Section 5.3, we concluded that even in the same collection, different query anchors can have varying requirements for the optimal multimodal fusion weights. This means that assigning a fixed weight for a group of query anchors can cause sub-optimal hyperlinking retrieval performance.

In Section 5.5, we applied an unsupervised learning solution using the Maximum Deviation Method (MDM) algorithm to optimise the linear fusion weights. The experimental investigation revealed that MDM can improve the multimodal fusion process. The runs combining the segment-level and video-level features achieved better hyperlinking results after using MDM to optimise the fusion weights.

In conclusion, this chapter contributes to our investigation of multimedia hyperlinking as follows:

- We proposed a set of investigations using multimodal information to improve hyperlinking performance. The experimental investigation indicates that the multimodal features, which are representative of cognitive information that users prefer when watching a certain video shot, benefit the hyperlinking performance.
- Using the linear fusion scheme, we can combine the hyperlinking results retrieved by multimodalities. We conclude that the optimal fusion weights for different query anchors might differ. Using the unsupervised solution (MDM) can provide a reliable optimisation of fusion weights.

There are also several issues which remain to be addressed. The experimental investigation raises a question of whether the segment-level visual features are necessary for hyperlinking retrieval. Our experiments demonstrated:

- Using only segment-level visual features caused decreased hyperlinking retrieval performance.
- The re-ranking strategy improved hyperlinking retrieval performance for the ME13data. However, applying this method to ME14data decreased hyperlinking retrieval.
- Fusing multiple segment-level features decreased MAP values compared with the baseline using only spoken information, even though these MAP values of fused results were greater than those using only equal weights.
- Fusing segment-level and video-level resulted in a slight improvement in terms of MAP. However, the experimental investigation failed to indicate which feature, the colour histograms or ORB descriptors, is a more effective contributor to the hyperlinking performance.

In general, the segment-level visual features used in this thesis exhibit unreliability in hyperlinking retrieval. We believed using equal fusion weights is one possible reason. However, after applying the MDM solution, the experimental results still showed a conflicting conclusion. During the discussion on the re-ranking strategy, we denoted that the initial retrieval can influence re-ranking performance. Moreover, the hyperlinking system should involve the retrieval model, identification of the target segment, and query anchor analysis. Until now, we assumed the spoken terms within the corresponding anchors to be input queries. In the next chapter, our investigation will focus on strategies to analyse the content of query anchors, with the aim of further improving multimedia hyperlinking performance.

Chapter 6

Improving Hyperlinking Performance by Query Anchor Analysis

6.1 Chapter Overview

The experimental investigations described in the previous chapter suggested that the effectiveness of multimodal features for video hyperlinking varies, and that these features can be complementary to each other. In Section 2.4.3, our review works concluded that combining multimodal features is one state-of-the-art approach in multimedia hyperlinking, and the other one is recreating query content. Thus, the experimental investigation in this chapter focuses on query anchor analysis to address the following research questions (RQ):

- RQ 5: How does recreating query anchor content improve hyperlinking retrieval?
- RQ 6: Can integrating query anchor recreation and multimodal features further improve hyperlinking results?

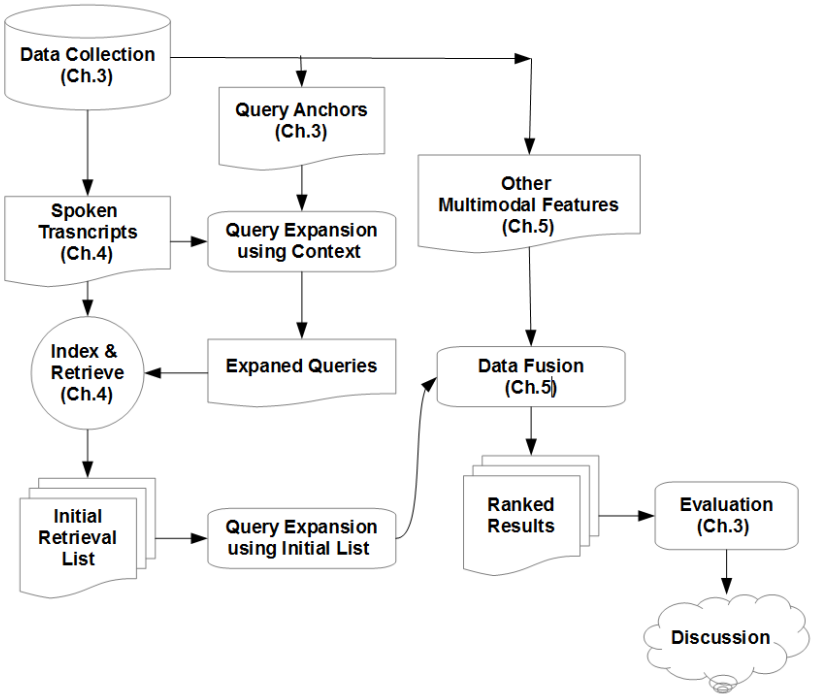


Figure 6.1: An overview of research design

The experiments in Chapter 4 and Chapter 5 used the spoken terms from the query anchor as the initial query. In this chapter, our research focuses on query anchor analysis to generate a richer description of the potential interesting content for content-based multimedia navigation (hyperlinking). A number of query expansion strategies applied to ASR transcripts are introduced and compared to investigate their potential for improving hyperlinking effectiveness. We integrate our query expansion methods with multimodal content analysis as described in Chapter 5. The experimental investigation shows how using query anchor analysis improves fusing multimodal features.

This chapter is structured as follows: Section 6.2 motivates and describes our investigation of the query expansion strategies based on spoken information. Our approaches are informed by existing MediaEval research and our previous investigation in TRECVid 2014; Section 6.3 presents experimental results of combining query anchor analysis and the multimodal fusion process. Experimental results

Table 6.1: Acronyms in experimental investigation

Abbreviation	Description
QE	query expansion method
Cxt	use query context in query expansion
RE	use pseudo ranked results and early fusion in query expansion
RL	use pseudo ranked results and late fusion in query expansion
E	use equal weights in late fusion
R	use rank-based normalisation in late fusion
H	use high-level concepts in the experiments

show that this strategy is superior to all other solutions proposed in this thesis in terms of MAP and P@N; Section compares our research conclusion with other research works proposed in MediaEval 2013, MediaEval 2014 and TRECVid 2015.

For readers' convenience, we propose Figure 6.1 and Table 6.1. Figure 6.1 illustrates a global view of experimental design proposed in this chapter, and Table 6.1 shows the acronyms of multimodal features and research methodologies. Reader can have a global view of our experimental design proposed in this chapter and a reference to check the acronyms in each experimental discussion.

6.2 Query Anchor Analysis

It was noted in [AKRR99] that a relevant document may fail to be retrieved if it does not contain the terms in the query, which emphasised the importance of query expansion for many IR tasks. Expanding query content aims to improve retrieval quality by reformulating an initial query to enrich its semantic information. There are several reasons to cause a discrepancy between an input query and potentially relevant documents in the database: users often lack sufficient experience to select appropriate terms to describe their information need; synonyms, morphological forms, or spelling errors increase the discrepancy between

Part I:
<top> <itemId>item_1</itemId> <queryText>Little Britian Fat Fighters problem of gypsies in the area</queryText> <visualQueues>fat club comedy</visualQueues> </top>
Part II:
item_1 v20080511_203000_bbcthree_little_britain 13.07 14.03

Figure 6.2: An example of searching query in MediaEval 2013 search subtask

what users have entered and what the IR system has acknowledged; it is possible that the document descriptions do not contain the terms within the query causing query-document term mismatch. These reasons have driven researchers to investigate methods to improve the representation of query content to encourage effective matching with relevant content.

A multimedia hyperlinking system is also confronted with the challenge of enriching the query content to best match potentially relevant segments, while, in some respects, this challenge is caused by the reasons mentioned above. In Chapter 3, we noted that hyperlinking queries contain no obvious input: a hyperlinking system needs to generate the query associated with the corresponding video segment. This is quite different from the query creation process in IR. The following gives two examples that illustrate the format of input queries in multimedia IR systems.

Figure 6.2 illustrates the input query for BBC TV data collection in the MediaEval Search task. The task requires participants to construct a retrieval system to search the TV data collection with the corresponding query. The query consists of two parts. The first part indicates the query ID (itemId) and a short piece of text describing what users are expecting when watching this segment. The second part contains a brief textual information to describe the visual cues (tagged as <visualQueues>).

Topic ID: 9099

Topic description: a checkerboard band on a police cap

Instance type: OBJECT

Visual example:

Figure 6.3: An example of the query in TRECVid 2014 Instance Search task

Figure 6.3 shows an example query in the TRECVid 2014 Instance Search Task (INS), which simulates the situation in which the user must find relevant video segments containing a certain instance (person, object, or place) within a large video collection. A topic description describes the search target, and the instance type (object or person) is provided. A further description of query content is illustrated by a set of keyframe examples. All the information can be used to create a multimodal interpretation for the query input to an instance search system.

Figure 6.4 presents an example query of the segment-based hyperlinking task in MediaEval 2014. The task applies to the same data collections (BBC TV data) as those presented in this thesis, and the TV data searching and hyperlinking task shares the same query boundary. The example consists of the video name and time interval for the query anchor. There is no text description or visual cues provided. Each participant must create query content using multimodal information.

The previous chapters have involved various strategies to generate the hyperlinking queries for our experimental investigations. In Chapter 4, the query

Part I:

```
<anchor>
<anchorId>anchor_1</anchorId>
<refId>53b3c46f42b47e459265d06f</refId>
<startTime>16.38</startTime>
<endTime>17.35</endTime>
<fileName>v20080629_184000_bbctwo_killer_whales_in_the</fileName>
</anchor>
```

Figure 6.4: An example of the hyperlinking query in MediaEval 2014 hyperlinking subtask

content was determined by all spoken terms detected by ASR algorithms. In Chapter 5, we used multimodal features to enrich the query content. The initial motivation for applying multimodal features was to enrich the content information in both target segments and query anchors to promote improved retrieval effectiveness. The experimental results demonstrated that low-level features can improve the hyperlinking performance by re-ranking the top retrieved results in ME13data. High-level features representing a summary of a video using textual (metadata) or visual (visual concepts) information can increase retrieval quality in terms of MAP for both the ME13data and ME14data test collections.

The query generation strategy used in Chapter 4 is extensible due to its simple implementation. Instead of using the spoken information directly, we expect that an enrichment of query content with potentially relevant spoken terms can increase the hyperlinking quality, and further improve multimodal hyperlinking. In Chapter 5, we showed that when applying the re-ranking strategy, a possible reason for decreasing results is the poor quality of hyperlinking results created by spoken information. In this section, we focus on expanding the hyperlinking query using ASR transcripts. The motivation of this research is to investigate efficient strategies for expanding hyperlinking queries using ASR transcripts in both ME13data and ME14data.

6.2.1 Query Expansion Strategy

To investigate the effectiveness of query expansion, we purpose two strategies based on spoken transcripts. The first is a simple method involving the context information of query anchors, which is inspired by [GP13]. [GP13] demonstrated that this method achieved best results in ME14data. The second applies pseudo relevance feedback to expand query content in our TRECVid 2014 Instance Search Task (INS) experiments [CMA+14]. In this section, we transplant this approach to the multimedia hyperlinking task, with an improvement to the data fusion scheme. The reasons of using these two methods are:

- We expect to investigate hyperlinking query expansion using both early fusion and late fusion mechanisms. The first method applies early fusion mechanism. It firstly implements query expansion then collecting retrieval results. While the second one applies late fusion mechanism by collecting retrieval results first and then implementing query expansion. Experimental investigation will compare the two approaches and discuss their effectiveness in multimedia hyperlinking systems.
- The method in [GP13] was proposed only for ME14data. We expect to investigate its effectiveness in ME13data.

Query Expansion using Context

The hyperlinking query involves no obvious user input to describe the potential searching requirement. The task of query expansion in a hyperlinking system is to enrich the description of a query segment to provide a more accurate representation of the cognitive information that could interest users. An ideal solution would be to use terms with high term frequency in the relevant documents and low frequency in the whole collection. In this experiment, we use unsupervised solution to detect the potential relevant terms associated with the corresponding

request since there is no reliable training set provided for either the ME13data or ME14data collections.

Robertson and Jones presented a simple and efficient query expansion strategy in [RJ94]. Its ability has been demonstrated in many subsequent studies [Ing96, Zha08, TTR12, EOS12]. The strategy is to calculate a weight, referred to as the "offer weight" according to [Rob90], for each potential expansion term related to the query after an initial retrieval run. Query expansion algorithms rank each term based on its offer weight and select the top ranked ones to modify query content. In [RJ94], the authors proposed that an offer weight for the ith term t_i in the query context is defined as Equation 6.1:

$$\text{OfferWeight}(t_i) = r_i \cdot \text{Score}_{\text{relevance}}(t_i), \tag{6.1}$$

where r_i indicates the number of relevant documents containing the term t_i. According to [RJ94], $\text{Score}_{\text{relevance}}(t_i)$ is defined as shown in Equation 6.2:

$$\text{Score}_{\text{relevance}}(t_i) = \log \frac{(r_i + 0.5) \cdot (N - n_i - R + r_i + 0.5)}{(n_i - r_i + 0.5) \cdot (R - r_i + 0.5)}, \tag{6.2}$$

where R is the number of known relevant documents for the current query, n_i is the number of documents containing the term t_i, N is the size of the document collection and r_i has the same definition as in Equation 6.1. [RJ94] used an experimental value 0.5 in Equation 6.2 to avoid division by zero in the absence of reference information and prevent removal of query terms no appearing so far in any relevant documents. Assume that we have a set of documents that are relevant to a query. We can extract all the terms within these documents and rank the terms according to their offer weights. Mathematically, a high offer weight means that this term has a low document frequency in the whole collection and higher document frequency in the relevant collection. Therefore, we can select the

top K terms to enrich the current query content. In [RJ94], the authors suggest that a reasonable value of K should be around 10 to 20.

Some issues in parameter setting need to be addressed before applying Robertson's method to enrich the hyperlinking query. The first one is how to determine the value of r_i. An IR system knows nothing about relevant documents when accepting a new query, meaning that r_i is unknown. To calculate the offer weight, we need to find a value of r_i. In a later section, the details associated with various methods are introduced. Another issue that we need to address is how to select the optimal value of K. The suggestion provided in [RJ94] could be unsuitable for hyperlinking retrieval. In our experimental investigation, we examine: 1) how changing the value of K influences the hyperlinking quality, and 2) whether we can indicate an optimal K for both ME13data and ME14data collections.

We hypothesise that an expanded query is based on the assumption: spoken information around the query anchor may be relevant and complementary to the query information and will be effective in improving the quality of the proposed hyperlinks. Thus, a temporal segment is defined before and after the query anchor, and the new query consists of all potentially relevant spoken terms detected by Robertson's query expansion method. The algorithm regards these terms as new query content to retrieve hyperlinks after removing stop words. Figure 6.5 illustrates this expansion process:

- Taking a hyperlinking query, we select all relevant terms from its context segments.
- Use Robertson's approach to select potentially relevant terms.
- Recreate the query by combining its content and those potentially relevant terms.
- Use the expanded query to retrieve hyperlinking results.

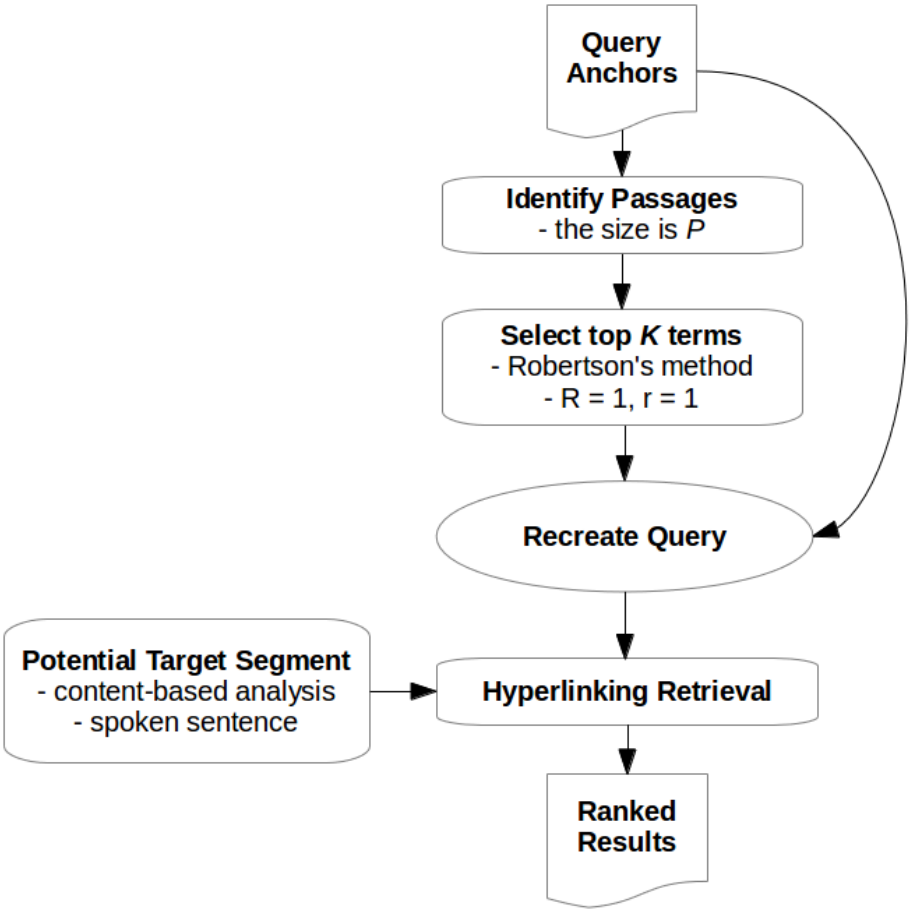


Figure 6.5: The workflow of using context information to expand query content

In Chapter 5, we demonstrated that there is no reliable training set for either ME13data or ME14data collections. Therefore, in this experiment, we use a grid search strategy to investigate: 1) the optimal size of temporal segments (P) and 2) the number of potentially relevant terms in query context (K). We calculate the offer weight of each term according to Robertson's method shown in Equation 6.1. We set R and r_i to be 1, meaning that there is only one potential relevant document (the relevant term set determined by the query context) in the collection. After ranking all the terms according to the offer weights, the top K terms are selected to be added to the original query content, and used as query anchors to retrieve the hyperlinking results.

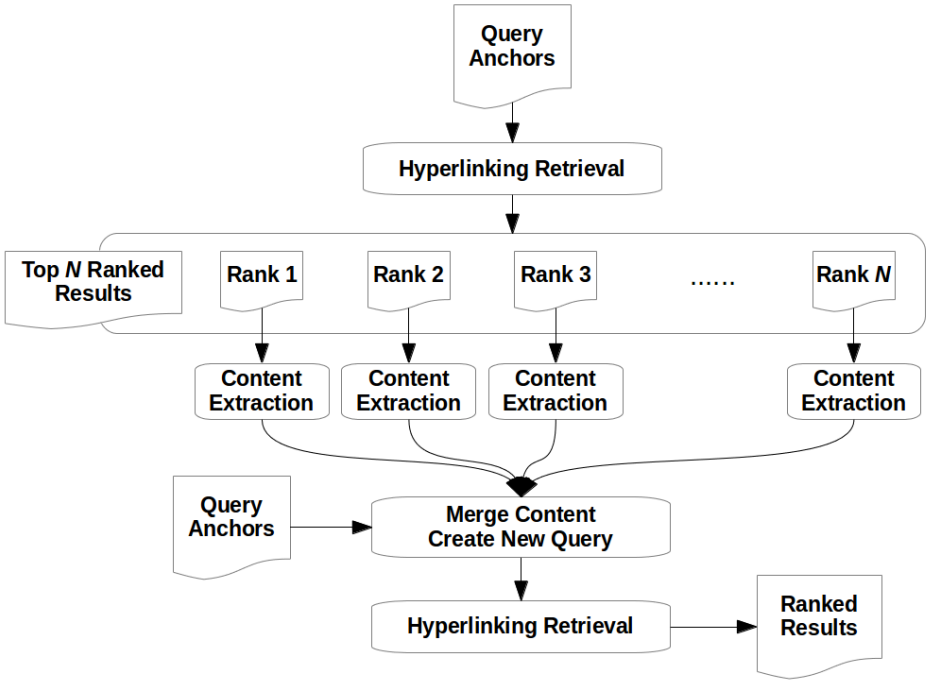


Figure 6.6: The workflow of using pseudo feedbacks and early fusion scheme

Query Expansion using Pseudo Relevance Feedback

Pseudo relevance feedback performs operates by performing an initial retrieval run after which the top R ranked documents are assumed to be relevant [GLJ11], and the final retrieval lists are created according to the information contained in these top R results. In a multimedia hyperlinking system, we assume that the top R initial results contain relevant terms that can be used to enrich the query. We use two methodologies to expand the query by applying pseudo relevance feedback.

Query Expansion using Early Fusion Our first method is to directly fuse the spoken terms at top R results retrieved using the initial query. This can be regarded as a variant of the context query expansion described previously, in which we replace the context segment with the top R results from the initial retrieval to extract potentially related terms. We use Equation 6.1 to calculate the offer weight of each term at top R results The value of r_i is the number of

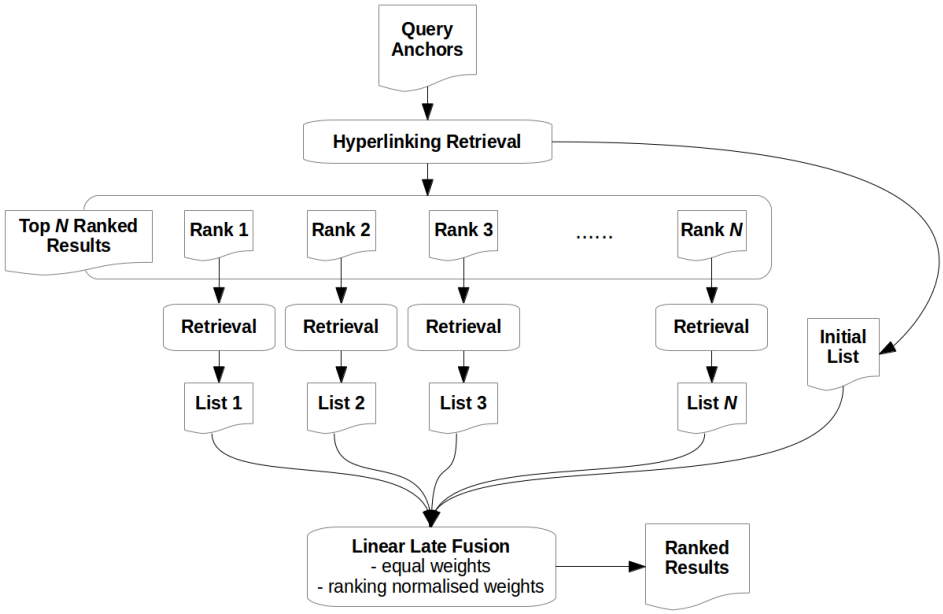


Figure 6.7: The workflow of using pseudo feedbacks and late fusion scheme

relevant documents containing the corresponding term within the top R results. After calculating the offer weights, the algorithm extracts the top K terms in each result to expand the query. The final input query includes spoken information from both the initial query anchor and a set of K terms. Figure 6.6 illustrates the workflow of the procedure.

Query Expansion using Late Fusion The second method is to apply a late fusion scheme to fuse initial results, and Figure 6.7 illustrates this process. An initial retrieval list, R_0, is created from the original query anchor. Overlapping segments are filtered from the retrieved ranked lists. Each result in the top R of the initial list is regarded as an individual pseudo query, and is used as a query to construct a new hyperlinking retrieval list, defined as R_i, where i means the ranking position of the pseudo query in R_0. The final score of each target segment seg is calculated by summing its score in the initial list and each pseudo list, as shown in Equation 2.1:

$$\text{Score}_{\text{seg}} = \sum_{i=0}^{N} w_i \cdot R_i(seg, q_i), \tag{6.3}$$

where w_i is the fusion score from each pseudo list. The function $R_i(seg, q_i)$ returns the score of seg in each retrieval list created by the query q_i where q_i is the retrieval result at rank i in the initial list R_0. The range of i is from 0 to R, where $i = 0$ indicates the initial list, and $i \in [1, ..., R]$ means the pseudo ranked list is retrieved using the ith segment for the top R initial list as the query input.

The value of the fusion weight w_i represents the proportion of terms relevant to the initial query in each ranked result. In our previous work [CMA+14], the fusion weight w_i was defined to be equal for each feature ($w_i = 1$). The evaluation results reported in [CMA+14] showed that using equal weights was not optimal. A simple assumption can be made that the higher the rank of the result, the more relevant terms it could involve. To estimate the fusion weights associated with a ranking, we apply the rank-based score methodology introduced in [Wu12]. This work proposed that the degree of relevance of a retrieved document can be determined by its ranking position. We assume that the level of relevance also indicates the number of terms that are potentially relevant to the query anchor in each ranked segment. Therefore, the fusing weight for each latent query is calculated according to Equation 2.7.

In conclusion, we use a total of four strategies to perform query anchor analysis for hyperlinking retrieval. The first two apply early fusion to implement query expansion by using Robertson's method. The second two applied late fusion to enrich the hyperlinking retrieval. All these approaches implement query expansion on spoken information. In the next section, we present a set of experiments to investigate and compare the hyperlinking quality for these expansion methods with various parameters.

6.2.2 Experimental Investigation

Query Expansion using Spoken Terms

This subsection focuses on investigating hyperlinking performance using the query expansion strategy: we apply Robertson's method to extract the potentially relevant terms from the context segments and pseudo retrieval results. To establish the optimal parameters, we set P (the size of context segment), K (the number of top ranked expansion terms used to enrich the query) and R (the top R results from the initial retrieval list) using the grid search strategy. For the two methodologies, Query Expansion using Context (QE-Cxt) and Query Expansion using Pseudo Ranked Results (QE-RE), the grid search stops when it is obvious that the hyperlinking results can not be improved.

Hyperlinking retrieval uses LIMSI transcripts. Our investigation compares the query expansion process with the baselines concluded in Chapter 4 (0.1633 for ME13data and 0.1528 for ME14data), which uses the same hyperlinking mechanism without applying query expansion. In this section, no visual features are used in hyperlink creation. Hyperlink construction follows the hypothesis described in Chapter 4, which uses BM25 to index and retrieve relevant segments associated with the parameters $b = 0.5$ and $k = 2.00$ for both data collections.

Table 6.2 shows hyperlinking performance using the strategy QE-Cxt on the ME13data collection, and Table 6.3 presents the results for the ME14 collection. The experiments confirm that using Robertson's query expansion strategy can improve hyperlinking performance. In Section 4.4, we showed that the MAP values for baselines using spoken information are 0.1633 for ME13data and 0.1524 for ME14data. All the MAP values in both tables outperform the corresponding baselines. Furthermore, the experiments suggest that the values of P and K should be moderate, when increasing P and K, the MAP value first increases,

Table 6.2: Hyperlinking results of ME13data (baseline: 0.1633) in terms of MAP using QE-Cxt (P stands for the size of segment, K means the number of merged terms).

P	60	120	180	240	300	360	420	480
K								
20	0.1914	0.1848	0.1861	0.1886	0.1853	0.1850	0.1849	0.1847
40	0.1970	0.1991	0.1888	0.1941	0.1860	0.1874	0.1852	0.1834
60	0.2015	0.2027	0.1926	0.1963	0.1914	0.1885	0.1845	0.1831
80	0.2012	0.2053	0.1978	0.1966	0.1961	0.1935	0.1886	0.1836
100	0.2015	0.2069	0.1942	0.2030	0.1970	0.1932	0.1927	0.1833
120	0.2023	0.2047	0.1965	0.2024	0.2002	0.1940	0.1964	0.1831
140	0.2022	0.2044	0.2011	0.2054	0.2007	0.1984	0.1977	0.1832
160	0.2025	0.2064	0.2029	0.2070	0.2033	0.1976	0.1976	0.1824
180	0.2019	0.2093	0.2042	0.2061	0.2037	0.2002	0.1987	0.1842
200	0.2019	**0.2105**	0.2060	0.2083	0.2048	0.1981	0.1971	0.1870
220	0.2019	0.2097	0.2036	0.2064	0.2074	0.1985	0.1973	0.1873
240	0.2019	0.2099	0.2037	0.2055	0.2066	0.1979	0.1956	0.1877
260	0.2019	0.2104	0.2035	0.2049	0.2029	0.1977	0.1944	0.1879
280	0.2019	0.2104	0.2019	0.2045	0.2016	0.1976	0.1939	0.1845
300	0.2019	0.2104	0.2012	0.2034	0.2009	0.1948	0.1931	0.1828

and then decreases. Apparently, using too many terms to enrich the query can introduce too much noise that is irrelevant to the hyperlinking request and lead to retrieval of non-relevant segments. We notice that when using a large P value, the experiments show an obvious decline in each level of the top K terms. On the other hand, choosing a large K also decreases MAP. For a smaller P (less than 180 seconds), we can observe that the MAP value decreases to a stable state when K reaches a particular value. This means that the algorithm has used all the terms within the current segment and increasing K can neither improve nor reduce the hyperlinking performance. For a larger P, we can also expect the MAP value to reach a stable state associated with a particular K. However, since there is an obvious drop of MAP values with a larger K (more than 200 seconds), we stop the experiments at $K = 300$.

The previous paragraph described a common result for using QE-Cxt with both collections. However, it is important to recognise the difference between the

Table 6.3: Hyperlinking results of ME14data (baseline: 0.1524) in terms of MAP using QE-Cxt (P stands for the size of segment, K means the number of merged terms).

P	60	120	180	240	300	360	420	480
K								
20	0.2159	0.2375	0.2428	0.2470	0.2473	0.2539	0.2539	0.2410
40	0.2186	0.2386	0.2471	0.2614	0.2688	0.2621	0.2641	0.2558
60	0.2203	0.2446	0.2575	0.2662	0.2652	0.2620	0.2636	0.2531
80	0.2193	0.2455	0.2641	0.2668	0.2646	0.2639	0.2598	0.2517
100	0.2145	0.2410	0.2559	0.2692	0.2737	0.2642	0.2574	0.2452
120	0.2141	0.2406	0.2549	0.2667	**0.2756**	0.2630	0.2574	0.2496
140	0.2141	0.2405	0.2547	0.2620	0.2722	0.2610	0.2532	0.2472
160	0.2141	0.2405	0.2541	0.2632	0.2615	0.2607	0.2535	0.2428
180	0.2141	0.2405	0.2546	0.2612	0.2651	0.2633	0.2525	0.2434
200	0.2141	0.2405	0.2545	0.2611	0.2645	0.2620	0.2552	0.2439
220	0.2141	0.2405	0.2545	0.2628	0.2654	0.2620	0.2502	0.2436
240	0.2141	0.2405	0.2545	0.2624	0.2648	0.2624	0.2502	0.2436
260	0.2141	0.2405	0.2545	0.2617	0.2648	0.2594	0.2508	0.2416
280	0.2141	0.2405	0.2545	0.2607	0.2644	0.2589	0.2501	0.2402
300	0.2141	0.2405	0.2545	0.2607	0.2642	0.2574	0.2489	0.2386

collections in these experiments. We cannot identify a unique parameter which achieves optimal results between the collections. For ME13data, the greatest MAP is achieved when assigning $P = 120$ and $K = 200$, whereas for ME14data, the optimal parameters are $P = 300$ and $K = 120$. This demonstrates that for ME14data, the context around query anchors contains more relevant information, and using a larger context size can benefit hyperlinking retrieval. This conclusion is also shown by comparing the greatest and lowest MAP values in the two tables to the baselines. For ME13data, the best improvement is 28.90% (from 0.1633 to 0.2105), and the lowest improvement is 12.11% (from 0.1633 to 0.1828), while for ME14data, the best is 80.84% (from 0.1524 to 0.2756), and the lowest is 40.49% (from 0.1524 to 0.2141).

From these experiments, we can conclude that:

Table 6.4: Hyperlinking results for ME13data (baseline: 0.1633) in terms of MAP using QE-RE (R means the number of pseudo relevant segments, and K means the number of merged terms from the corresponding segment)

R	5	10	15	20	25	30	35	40
K								
5	0.1725	0.1625	0.1614	0.1503	0.1382	0.1347	0.1311	0.1255
10	0.1795	0.1616	0.1638	0.1550	0.1422	0.1357	0.1302	0.1243
15	**0.1828**	0.1636	0.1582	0.1544	0.1415	0.1361	0.1293	0.1260
20	0.1824	0.1645	0.1619	0.1525	0.1407	0.1350	0.1328	0.1221
25	0.1807	0.1680	0.1616	0.1491	0.1393	0.1364	0.1287	0.1208
30	0.1813	0.1664	0.1633	0.1490	0.1397	0.1331	0.1297	0.1238
35	0.1819	0.1660	0.1604	0.1515	0.1376	0.1332	0.1304	0.1239
40	0.1815	0.1643	0.1605	0.1478	0.1416	0.1336	0.1314	0.1270

- Query expansion using context information can significantly improve hyperlinking performance on spoken information. The results for both collections show a significant improvement in terms of MAP.

- Assigning moderate parameters is critical to achieving an optimal MAP value. Large P or K values can result in lower hyperlinking quality.

- We observe that the optimal parameters within the two collections are different, the major reason being that the context information in ME14data contains more relevant terms than that in ME13data.

Next, we plan to compare this strategy (QE-Cxt) with the other query expansion strategy (QE-RE).

Tables 6.4 and 6.5 present hyperlinking results using QE-RE in terms of MAP metrics. We observe the best MAP value in Table 6.4 when $R = 5$ and $K = 15$, and in Table 6.5 when $R = 5$ and $K = 30$. A large increment of R and K can cause a significant decline in MAP values. The experiments demonstrate that selecting the top 5 pseudo retrieval results ($R = 5$) is optimal, and that, with an appropriate range of pseudo feedback, the QE-RE strategy can achieve better performance

Table 6.5: Hyperlinking results for ME14data (baseline: 0.1524) in terms of MAP using QE-RE (R means the number of pseudo relevant segments, and K means the number of merged terms from the corresponding segment)

				TOP	Rank			
R	5	10	15	20	25	30	35	40
K								
5	0.2017	0.1990	0.2068	0.2040	0.1978	0.1972	0.1947	0.1895
10	0.2145	0.2148	0.2111	0.2063	0.2053	0.2047	0.2014	0.1975
15	0.2217	0.2166	0.2107	0.2060	0.2084	0.2077	0.2058	0.1985
20	0.2261	0.2141	0.2090	0.2059	0.2102	0.2065	0.2069	0.2005
25	0.2295	0.2154	0.2109	0.2052	0.2091	0.2054	0.2042	0.1978
30	**0.2321**	0.2170	0.2095	0.2060	0.2102	0.2055	0.2029	0.1918
35	0.2320	0.2163	0.2113	0.2078	0.2113	0.2069	0.1991	0.1892
40	0.2308	0.2183	0.2131	0.2105	0.2095	0.2028	0.1929	0.1938

to the baselines. When increasing the value of R, we can observe a significant drop in MAP values for the ME13data collection. When R exceeds 20, the MAP values in Table 6.4 are lower than the corresponding baseline (0.1633). Meanwhile, we notice that the MAP decline in Table 6.5 is small. The lowest MAP (0.1895) is still higher than the corresponding baseline (0.1524). We conclude that QE-RE can improve hyperlinking performance with proper parameter assignment. An inappropriate parameter assignment can produce a worse hyperlinking result than the baseline, as demonstrated in Table 6.4.

The experiments do not agree on an optimal K for both collections. In [RJ94], the author suggests a safe range of K from 10 to 20. In one aspect, we confirm this conclusion in our experiment on the BBC TV collection, with the evidence that when $K = [10, 20]$, the corresponding MAP values are better than the baselines. Moreover, in Table 6.5, the greatest MAP value occurs when $K = 30$, which suggests that the reasonable range of K in our hyperlinking system can be $[10, 30]$.

Comparing QE-Cxt and QE-RE, we conclude that:

- Context information is critical to enrich the query content. Experiments for both data collections showed that using QE-Cxt can achieve better hyperlinking performance in terms of MAP than QE-RE. For ME13data, we

concluded that the best improvement for QE-Cxt was 28.90%, but that the rate for QE-RE is only 11.92% (from 0.1633 to 0.1828). For ME14data, the improving rate for QE-Cxt is 80.84%, which is much higher than a 52.30% rate for QE-RE .

- The query expansion strategy is more efficient for the ME14data collection. All the experiments demonstrate that the improvement in ME14data are higher than that in ME13data using either of the strategies. In Figure 5.3, we showed that in ME14data using queries without any expansion caused extremely low retrieval results in some cases. We believe that applying query expansion enriches those queries and achieves better improvement in ME14data.

- QE-Cxt is superior to QE-RE in both test collections. The first evidence is that in both collections, the best MAP value achieved by QE-Cxt is higher than that using QE-RE. We showed that all the MAP values achieved by QE-Cxt are superior to the corresponding baselines in Table 6.2 and Table 6.3, while in Table 6.4, when assigning improper parameters, QE-RE receives a lower MAP value than the corresponding baselines.

Query Expansion using Late Fusion

This section investigates strategies using a late fusion scheme to analyse the query anchor for multimedia hyperlinking. The methodologies are:

- **Query Expansion using Late Fusion Scheme with Equal Fusion Weights (QE-RL-E)** The methodology is presented in Section 5.2.1. We apply equal fusion weights to implement the late fusion process on the ranked lists retrieved by the initial query.

Table 6.6: Hyperlinking results for both ME13data (baseline: 0.1633) and ME14data (baseline: 0.1524) in terms of MAP using QE-RL

	R=5	R=10	R=15
ME13data			
QE-RL-E	0.1752	0.1749	0.1744
QE-RL-R	**0.1806**	0.1778	0.1768
ME14data			
QE-RL-E	0.2125	0.2010	0.1984
QE-RL-R	**0.2205**	0.2130	0.2079

- **Query Expansion using Late Fusion Scheme with Ranked Normalised Scores (QE-RL-R)** The methodology is presented in Section 5.2.1. We apply the ranked normalised scores (Equation 6.3) to fuse the ranked lists retrieved by the initial query.

The experiment examines both strategies on ME13data and ME14data collections. The initial query is determined according to the process presented in Chapter 4, as using all the spoken terms within the query boundary without any further expansion. The hyperlinking process follows the hypothesis introduced in Chapter 4.1.

Table 6.6 presents the hyperlinking results for QE-RL-E and QE-RL-R in both collections in terms of MAP. We notice that both methods can outperform the corresponding baselines. When $R = 5$, both methods achieve their greatest MAP values. For ME13data, the best improvement is 10.59% (from 0.1633 to 0.1806), and for ME14data, the best improvement is 44.68% (from 0.1524 to 0.2205). The experiments confirm that using the ranked normalised scores can improve hyperlinking performance, with the evidence that the MAP of each QE-RL-R run is higher than that of the corresponding QE-RL-E run. The linked segment with a lower rank is less relevant to the initial query, and the fusing weights for the

ranked list retrieved by these segments should be lower to decrease the influence of irrelevant information. The conclusion is also supported by the clear decline of MAP when increasing R.

6.2.3 Discussion

In this section, we presented a set of strategies to enrich hyperlinking queries to improve hyperlinking performance. We divided these methodologies into two categories: expanding the query content using spoken terms, and using pseudo retrieval results and a data fusion scheme to enrich the final hyperlinking results. The former used early fusion to enrich the query content, and the strategies were represented by the abbreviations QE-Cxt and QE-RE. The latter used late fusion to recreate the hyperlinking results, and the strategies were referred to as QE-RL-E and QE-RL-R.

The experiments revealed that all the methodologies outperformed the baseline concluded in Chapter 4. We concluded that further processing on the hyperlinking query could improve hyperlinking performance. We note that the improvement should be with proper estimation of parameters in each method. All the methods using pseudo retrieval results (QE-RE, QE-RL-E, and QE-RL-R) achieve the best hyperlinking quality when using the top 5 pseudo feedbacks ($R = 5$). Increasing the number of pseudo feedback items decreases performance. For QE-Cxt, we could not identify parameters (K and P) that would be optimal for both our experimental datasets. The experiments suggest that the parameters should be moderate since setting low or high those parameters decreases hyperlinking results in terms of MAP.

Table 6.7 illustrated the improvement between the baselines and all the best results for each method. We concluded that QE-Cxt is the best method to enrich query content. In Chapter 5, we noted the poor quality of the initial query

Table 6.7: An analysis of the improvement in MAP values among various strategies for query expansion

Baseline	QE-Cxt	QE-RE	QE-RL-E	QE-RL-E
ME13data 0.1633	0.2105/28.90%	0.1828/11.94%	0.1752/07.29%	0.1806/10.59%
ME14data 0.1524	0.2756/80.84%	0.2321/52.30%	0.2125/39.44%	0.2205/44.68%

for ME14data, which caused a low MAP for some hyperlinking requests. We believed that this is the primary reason for the better performance of query anchor expansion in ME14data compared to ME13data.

The great MAP achieved by QE-Cxt implies that video segments around a query anchor contain significant relevant information. However, it is important to be aware that direct expansion of the query anchor could increase the number of retrieved segments located within the same video, as those potential segments within a segment could get a better match with the expanded query content. To investigate this, we use the average number of relevant segments of the query video, defined as PV@R in Equation 6.4:

$$\text{PV@R} = \frac{\sum_{j=0}^{M} \text{Relevant}(R_j, K)}{M}, \tag{6.4}$$

where M is the number of input queries, R_j means the retrieved list created for the jth query, and the function Relevant returns the number of relevant segments within the video containing the corresponding query anchor for the top K results in R_j. Table 6.8 shows PV@R for all strategies where R is set to be 5, 10, and 20.

Table 6.8 demonstrates that QE-Cxt has the advantage of retrieving more relevant segments within the query video. In general, QE-RL-R retrieves the least number of segments within the query video. However, the difference is

Table 6.8: PV@R values of all query expansion strategies for ME13data and ME14data

PV@RF	QE-Cxt	QE-RE	QE-RL-E	QE-RL-R
ME13data				
R=5	**2.0476**	1.3809	1.3333	1.1428
R=10	**2.7619**	1.8571	1.8571	1.5238
R=20	**3.3809**	2.1428	2.1904	2.0476
ME14data				
R=5	**3.1000**	1.4667	1.4333	1.4333
R=10	**4.5667**	2.2667	2.3000	2.3333
R=20	**5.7333**	3.5000	3.2333	3.3000

small compared with the other two methods (QE-RE and QE-RL-E). We conclude that when using early fusion, the context information contains more relevant information for hyperlink construction, compared with the segments within the pseudo feedback results. The QE-RE strategy is based on the assumption that the top R results in the initial retrieval are related to the linking topic. However, the lower MAP value of QE-RE indicates the negative influence of irrelevant segments within the top R results.

The experimental results show that QE-RL always receives a lower MAP value than those using QE-Cxt and QE-RE. However, QE-RL can use the query generated by QE-Cxt and QE-RE to create hyperlinks, which implies the possibility of improving hyperlinking performance by combining these two strategies. This means that the combination could inherit the advantage of both methodologies. We have already demonstrated that QE-Cxt outperforms on other methods at retrieving the relevant segments within the query video. In the next section, we use multimodal feature analysis in an attempt to further improve the QE-RL strategy.

6.3 Combine Query Anchor Analysis with Multimodal Features

Chapter 5 proposed a hyperlinking model using video-level features to improve hyperlinking performance using video-level features. Using the MDM algorithm, we concluded that fusing video-level features can significantly increase retrieval quality. In the previous section, we applied query anchor analysis to improve hyperlinking retrieval. Experimental investigation demonstrated that query anchor analysis can outperform the baselines concluded in Chapter 4. Using context spoken information around the query anchor was shown in these experiments to be an effective approach. In this section, we combine multimodal features with the query anchor analysis. Our motivation is: 1) to investigate whether video-level features always benefit hyperlinking retrieval associated with various query expansion strategies; 2) which query expansion method performs better when fusing video-level features; 3) whether applying low-level visual features can improve hyperlinking retrieval with expanded queries.

6.3.1 Using the Video-level Features

In this section, the query anchors to retrieve multimedia hyperlinks are created according to the approaches mentioned in the previous section:

- Expand query anchors using time-based context (QE-Cxt). The method to extract spoken information follows the procedure described in Section 5.1. Based on the previous experiments, the optimal parameters for ME13data are $P = 120$ and $K = 200$, and for ME14data $P = 300$ and $K = 120$.
- Recreate the content of the query anchor using the top R initial retrieval (QE-RE) results. The initial retrieval is created according to the description in Section 6.2. We define the value of R to be 5.

- Define the top R results in the initial retrieval results as potential query anchors and collect R new hyperlinking results using the corresponding query anchor. Then we have R lists retrieved by potential query anchors, and 1 list retrieved by the initial query anchor. A late fusion scheme is applied to fuse these $R + 1$ retrieval lists. We demonstrated that the optimal choice to determine the fusion weight is by using ranked-based normalised scores. However, considering the uncertainty of hyperlinking performance when applied to multimodal fusion, we decide to explore both presented solutions: equal fusion weights (QE-RL-E) and rank-based normalised weights (QE-RL-R). The value of R is set to be 5 according to the previous experiments.

Both video-level features, including the concept feature (CPT) and metadata (META), are used to implement hierarchy hyperlinking. The methodologies to retrieve video-level features and implement linear fusion were proposed in Chapter 5. Spoken information (LIMSI transcripts) indexing and search follow the procedure proposed in Chapter 4. Fusing multimodal features is implemented using linear combination as defined in Equation 5.10, with the MDM algorithm determining the fusion weights. All the presented methodologies are applied to both ME13data and ME14data collections. We define the baseline as multimodal fusion hyperlinking retrieval using the MDM algorithm. The evaluation benchmarks involve both MAP and P@5[1]

Figures 6.8 and 6.9 illustrate the evaluation results in terms of MAP when applying the combined solution to ME13data and ME13data. The results demonstrate that the combined solution outperforms the baselines. The experiments reveal that META is the optimal video-level feature, with the evidence that all the runs using CPT receive lower results than with those using only META.

[1]The evaluation benchmark provided by MediaEval organizers indicates P@5, 10, 20 without P@1. Thus, we select P@5 as the primary metrics since a linking system should link resources to most relevant entities.

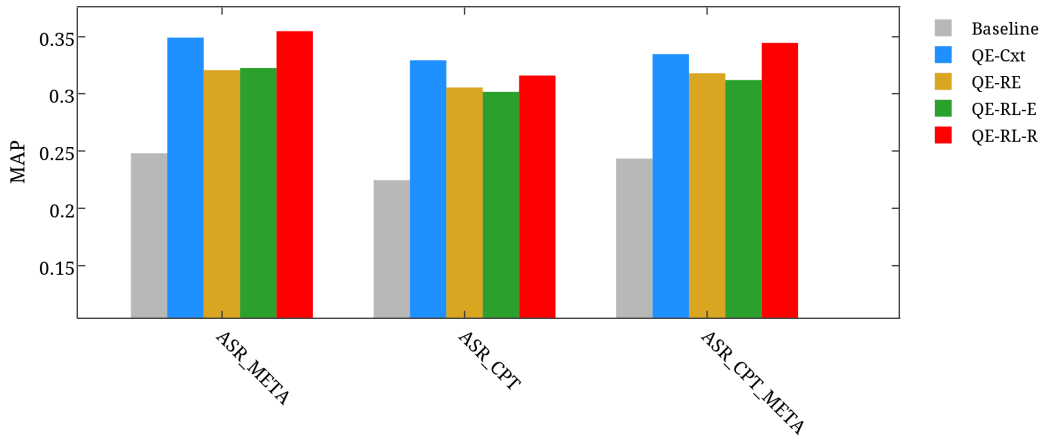


Figure 6.8: Hyperlinking results for combining the video-level feature (META or CPT) with the query expansion strategies for ME13data (MAP)

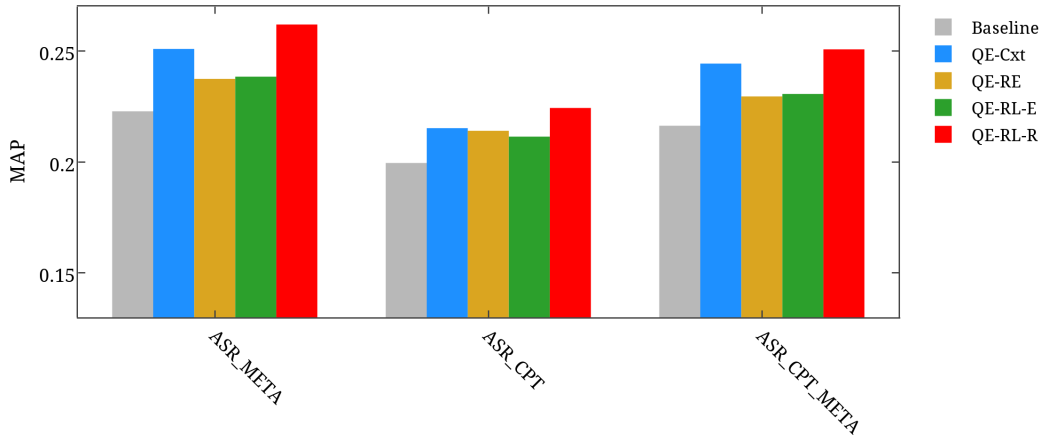


Figure 6.9: Hyperlinking results for combining the video-level feature (META or CPT) with the query expansion strategies for ME14data (MAP)

Table 6.9: An analysis of MAP values when combining spoken information (ASR) with various video-level features (META or CPT)

	ASR	ASR_META	ASR_CPT	ASR_CPT_META
ME13data				
QE-Cxt	0.2105/-	0.2509/19.19%	0.2152/02.23%	0.2443/16.06%
QE-RL-R	0.1806/-	0.2619/45.02%	0.2243/24.20%	0.2507/38.82%
ME14data				
QE-Cxt	0.2756/-	0.3491/26.67%	0.3293/19.48%	0.3347/21.44%
QE-RL-R	0.2205/-	0.3547/60.86%	0.3160/43.31%	0.3445/56.24%

Comparing the two methods to expand query content (QE-Cxt and QE-RE), the results in both figures confirm that QE-Cxt is a better solution, given the evidence of its superior MAP values. Comparing the two methods to pseudo retrieval (QE-RL-E and QE-RL-R), we conclude that QE-RL-R performs better for combination with video-level features in terms of MAP. Considering the previous experiments, we can confirm that QE-Cxt and QE-RL-R can improve hyperlinking query analysis. Therefore, in the remainder of this section, our investigation focuses on these two approaches.

In Figure 6.9, we observe that when fusing CPT, the MAP for QE-RL-R is lower than that for QE-Cxt. However, in all other cases, QE-RL-R always receives a higher MAP than QE-Cxt. The experiments in the previous section received the superiority of QE-Cxt using only spoken information. This conclusion has to be revised when considering the influence of video-level features. Table 6.9 shows the improvement in MAP using only spoken information to enrich query content, and combining the query analysis with video-level features. The results demonstrate that QE-RL-R achieves a better improvement in all cases.

The QE-Cxt and QE-RL-R strategies represent two methodologies to analyse a hyperlinking query. The former uses the early fusion scheme, and the latter uses the late fusion scheme. Technically, the hyperlinking query created by QE-Cxt can be used by QE-RL-R, which suggests a combination of these two methods could

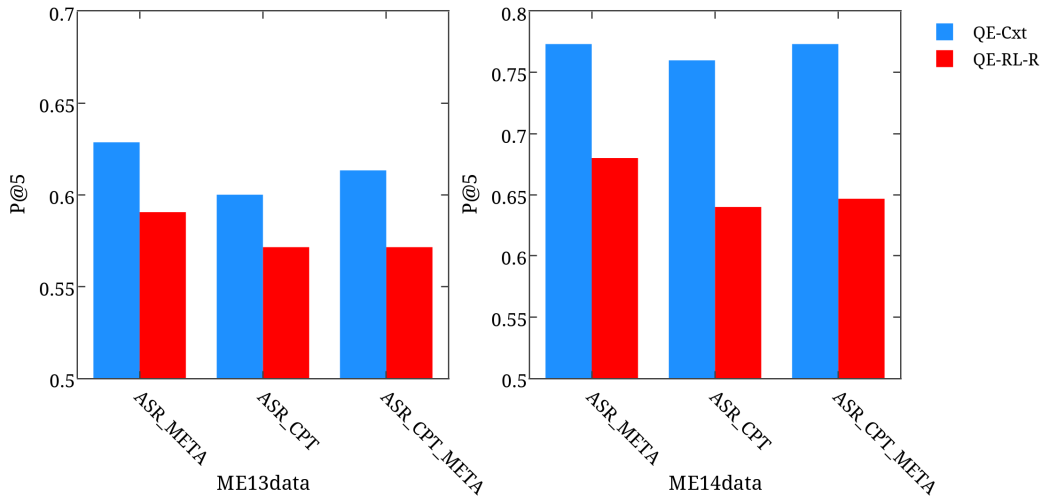


Figure 6.10: A comparison of P@5 between QE-Cxt and QE-RL-R for ME13data and ME14data

improve hyperlinking performance. The previous experiments demonstrated the superiority of QE-RL-R in terms of MAP. We have outlined the advantage of QE-Cxt in retrieving segments within the query video in Section 6.1, which results in a higher precision rate at top R results. In Figure 6.10, we compare the corresponding P@5 values of QE-Cxt and QE-RL-R when fused with video-level features.

Figure 6.10 reveals that although QE-RL-R outperforms other strategies in terms of MAP, the QE-Cxt achieved a better performance in terms of P@5. The best P@5 in ME13data is 0.6286 at ASR_META, and the best P@5 in ME14data is 0.7733 using both ASR_META and ASR_CPT_META. The P@5 value of ASR_CPT (0.7600) is slightly lower than the best result (0.7733). According to the previous experiments, we conclude that the QE-Cxt strategy is superior at retrieving hyperlinks with high precision rate, and that the QE-RL-R strategy is excellent at improving the recall rate. Before, we noted that the MAP of QE-RL-R is lower than that of QE-Cxt. Thus, the improvement of MAP values using QE-RL-R is achieved by using video-level features, especially metadata information.

In conclusion, the experiments confirm the importance of multimodal feature analysis for improving hyperlinking quality. Compared with the two baselines, one is using video-level features, and the other is using query anchor analysis, the combined methodology can significantly increase hyperlinking performance in terms of MAP. The experimental results support the conclusion proposed in Chapter 5, that the metadata information is more effective than high-level concepts when representing video-level information. In the next section, we propose a combination of QE-Cxt and QE-RL-R to create a better overall hyperlinking retrieval framework.

6.3.2 An Integrated Framework for Multimedia Hyperlinking

This section proposes an enhanced hyperlinking framework which combines the query anchor analysis and hierarchy hyperlinking models. In general, the framework uses the two presented methods to enrich hyperlinking queries, including QE-Cxt and QE-RL-R. The hyperlinking framework accepts the refined query to retrieve relevant segments as the initial retrieval. Video-level multimodal information is then applied to improve the initial retrieval to create the final linked results. The workflow of the enhanced framework is shown in Figure 6.11.

In the proposed framework, the hyperlinking query is enriched by merging the context information. The merged spoken terms are determined by Robertson's algorithm as proposed in Section 6.1. The framework accepts the merged query and uses the initial hyperlinking retrieval list based on spoken information (ASR transcripts created by LIMSI algorithm). Then, the top R results in the initial retrieval list are extracted and taken as queries. The framework then creates R new hyperlinking results according to these queries. The late fusion scheme is applied to merge the $R + 1$ lists. The fusion weights are determined by ranked normalised scores. Finally, the hyperlinking result using spoken information is

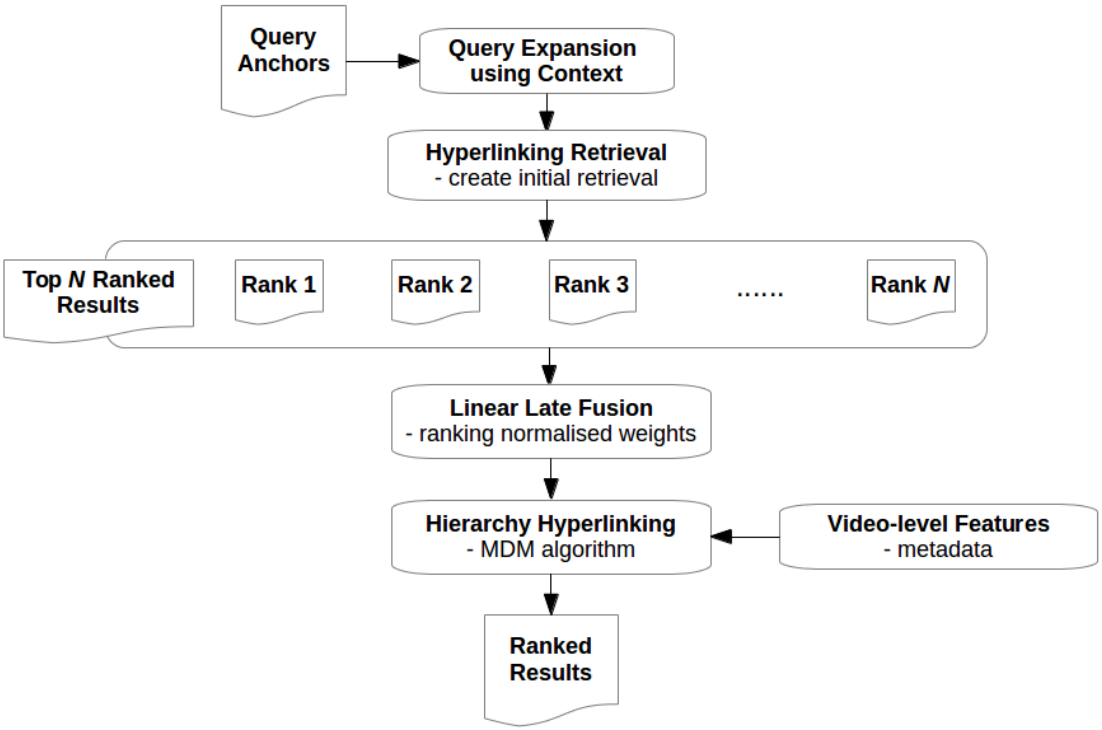


Figure 6.11: A multimedia hyperlinking model combining query anchor analysis and the hierarchy hyperlinking strategy

fused with video-level features. The MDM algorithm described in Chapter 5 is used to estimate the fusion weights between segment-level and video-level features. In the following experiment, we use the abbreviation H-QE-Cxt-RL to represent this framework.

To investigate the proposed hyperlinking framework, we select the best strategy to implement each step of the multimodal process. When expanding the query content, we use the optimal parameters in the previous experiment for each collection. For ME13data, we use $P = 120$ and $K = 200$, and for ME14data we use $P = 300$ and $K = 120$. Before investigating H-QE-Cxt-RL, we need to determine the top R results to create pseudo retrieval lists. In the previous section, we concluded that for QE-RE and QE-RL, the optimal value of R is 5. However, there is no evidence that this conclusion can be applied to our new hyperlinking framework. Therefore, we first investigate the optimal R for H-QE-Cxt-RL.

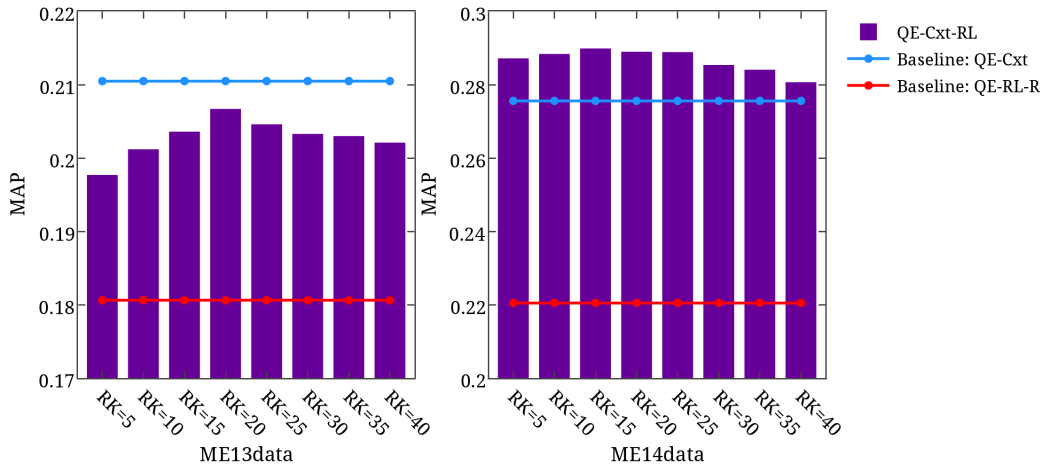


Figure 6.12: Hyperlinking results in terms of MAP using QE-Cxt-RL for ME13data and ME14data

As shown in Figure 6.11, the linear fusion step occurs before fusing the video-level features. We first examine the value of R without the integration of the video-level features. This strategy is referred to as QE-Cxt-RL, which means we only use the QE-Cxt and QE-RL schemes to implement hyperlinking retrieval based on spoken information. We investigate values of R from 5 to 40.

Figure 6.12 shows hyperlinking performance for QE-Cxt-RL in terms of MAP. The results reveal that the combined QE-Cxt-RL strategy can produce higher MAP than QE-RL. However, for ME13data, the hyperlinking quality is lower than QE-Cxt. For ME14data, we observe that QE-Cxt-RL is superior to the other two baselines. Considering the experiments presented in the previous section, we confirm that using only QE-RL can not produce satisfactory results.

The primary task of QE-Cxt-RL is to indicate the optimal R for H-QE-Cxt-RL. Figure 6.12 shows that a reasonable range of R is from 15 to 20. The best MAP occurs when fusing the top 20 results for ME13data, and for ME14data, it occurs when the R is set to 15. We noted that the experimental investigations shown in Tables 6.4 and 6.5 indicate that the optimal R is 5 when fusing the

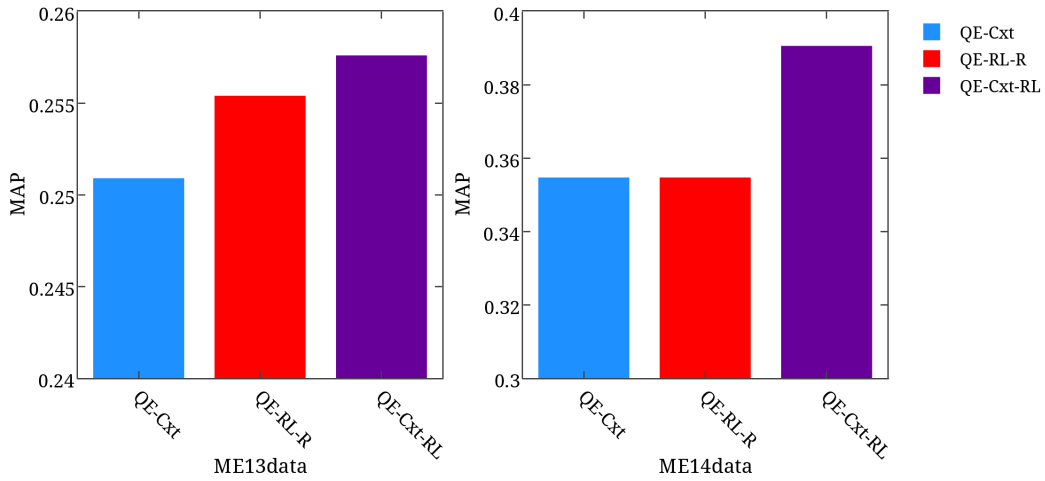


Figure 6.13: A comparison of H-QE-Cxt, H-QE-RL and H-QE-Cxt-RL in terms of MAP, ME13data and ME14data

initial retrieval list, while the experiments proposed in Figure 6.12 suggest that when combining QE-Cxt and QE-RL, we need to increase R to achieve better hyperlinking performance. Previously, the initial list directly retrieved by spoken information had a low $P@N$ value. This means that increasing R could introduce more irrelevant results and shift the focus of the query content. The current approach creates the initial hyperlinking list using the QE-Cxt strategy, which has been demonstrated to be effective at increasing $P@N$ values. The solution involves more relevant segments from the top results in the initial retrieval list as the expansion queries. Thus, in Figure 6.12, we can observe that the optimal value of R increases to 20 in ME13data and 15 in ME14data.

To investigate the effectiveness of the value of H-QE-Cxt-RL hyperlinking model, we compare it against two baselines:

- QE-Cxt is used to create the hyperlinking query. The hyperlinking results are re-ranked by fusing the video-level features (metadata) according to the strategy proposed in Section 5.3, and the fusion weights are determined by

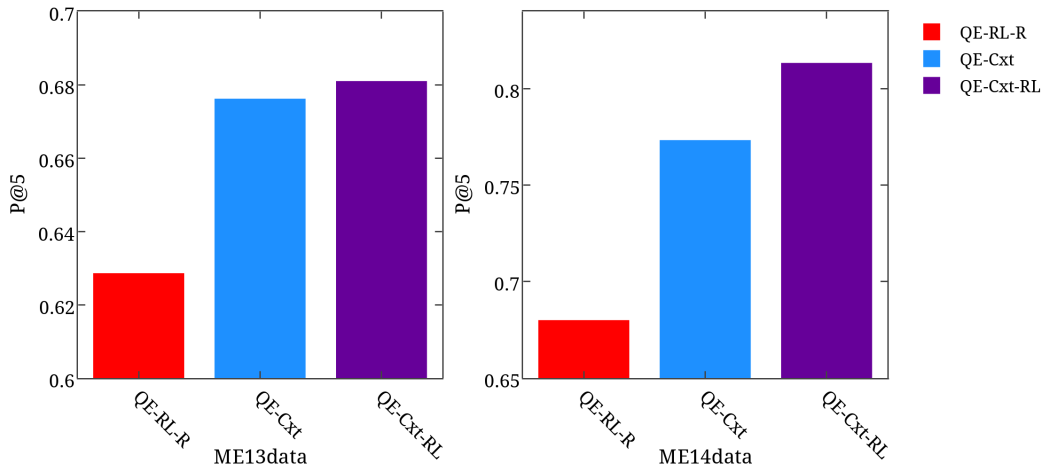


Figure 6.14: A comparison of H-QE-Cxt, H-QE-RL and H-QE-Cxt-RL in terms of P@5, ME13data and ME14data

using the MDM algorithm proposed in Section 5.5. The resulting MAP value is illustrated in Figure 6.8. We use the abbreviation H-QE-Cxt to represent this approach in the later discussion.

- QE-RL-R is used to create the hyperlinking query. The hyperlinking results are re-ranked by fusing the video-level features (metadata) according to the strategy proposed in Section 5.3, and the fusion weights are determined by using the MDM algorithm proposed in Section 5.5. The corresponding MAP value is illustrated in Figure 6.8. We use the abbreviation H-QE-RL to represent this approach in the later discussion.

For H-QE-Cxt-RL we use the best parameters as identified in Figure 6.12. This means that, for ME13data, we fuse the top 20 results from the initial retrieval, and for ME14data, we fuse the top 15. Figure 6.13 shows the resulting MAP values, while Figure 6.14 plots the corresponding P@5 values .

These experiments demonstrate that when using the hierarchy hyperlinking model, the strategy H-QE-Cxt-RL strategy outperforms the H-QE-Cxt and H-QE-RL methods in terms of both MAP and P@5. This demonstrates that a

combined methodology of query anchor analysis and hierarchy hyperlinking model can improve hyperlinking performance. We conjecture that the QE-Cxt scheme favours a high precision rate for the initial hyperlinking list, which is demonstrated by varying R in Figure 6.12, while the QE-RL strategy performs in a complementary way to increase the number of relevant segments with relatively lower ranks. However, to locate these relevant segments at a higher rank, the fusion of video-level features is essential. It has been demonstrated that QE-RL achieves better performance after being fused with the metadata information, shown in Figure 6.8 and Figure 6.9. We can observe the same conclusion by comparing the hyperlinking results illustrated in Figure 6.12 and Figure 6.13. Figure 6.12 demonstrates that for ME13data, the QE-Cxt-RL strategy, without a fusion of video-level features, was less effective than QE-Cxt. In Figure 6.13, the results using pseudo feedbacks, H-QE-RL and H-QE-Cxt-RL, produce higher MAP values H-QE-Cxt.

In conclusion, we have proposed our hyperlinking framework shown in Figure 6.11. Its effectiveness has been demonstrated in the previous experimental investigation. In the next section, we address the unsolved question regarding segment-level features raised in Chapter 5.

6.3.3 An Investigation to Segment-level Features

In Chapter 5, we classified the multimodal features in a hyperlinking system as segment-level and video-level. Both of these can be used to complement the information in a hyperlinking query, as illustrated in Section 5.5 and Section 6.3.1. The previous experiments in Chapter 5 and this chapter demonstrated that video-level features can improve hyperlinking performance. However, experiments in Section 5.2.2 showed that segment-level features, HSV colour histogram (CH) and ORB descriptor (ORB), decreased hyperlinking retrieval. Furthermore, we

Table 6.10: An analysis of MAP results for fusion of transcripts with segment-level features.

	Baseline	CH	ORB	CH+ORB
ME13data				
QE-Cxt	0.2105	0.1944/**-03.52**%	0.1723/**-17.72**%	0.1815/**-13.78**%
QE-RL	0.1806	0.1563/**-13.46**%	0.1485/**-17.77**%	0.1525/**-15.56**%
QE-Cxt-RL	0.2067	0.1618/**-21.72**%	0.1435/**-30.58**%	0.1528/**-26.08**%
ME14data				
QE-Cxt	0.2756	0.2718/**-01.38**%	0.2635/**-04.39**%	0.2646/**-03.99**%
QE-RL	0.2205	0.2141/**-02.90**%	0.1967/**-10.79**%	0.1975/**-10.43**%
QE-Cxt-RL	0.2889	0.2801/**-03.05**%	0.2759/**-04.50**%	0.2782/**-03.70**%

proposed the re-ranking strategy using late fusion to integrate segment-level features. Based on the experimental results shown in Figure 5.2, we can conclude that the re-ranking strategy failed to increase the hyperlinking results due to the poor quality of the initial results retrieved using spoken information. From Figure 5.13, we can observe that late fusion between segment-level features and spoken information can achieve improved MAP in some cases. We note, however, that the hyperlinking results using video-level features are superior to those using only segment-level features. All these experiments lead us to reconsider the value of visual segment-level features: whether those features can make a stable contribution to improving hyperlinking quality. In this section, we integrate segment-level feature analysis to the current approach. The methodologies include re-ranking the top R results and retrieving hyperlinking results using late fusion:

- Use late fusion to integrate the segment-level features with other multimodal features, which involves both spoken information and video-level features.
- Re-rank the top R results of initial retrieval lists, which are constructed by using query anchor analysis and the hierarchy hyperlinking model.

Table 6.11: An analysis of MAP results for fusion of transcripts with segment-level features and video-level features

	META	META+CH	META+ORB	META+CH+ORB
ME13data				
QE-Cxt	0.2509	0.2427/**-03.27**%	0.2338/**-06.82**%	0.2283/**-09.01**%
QE-RL	0.2554	0.2315/**-09.58**%	0.2468/**-03.37**%	0.2311/**-09.51**%
QE-Cxt-RL	0.2576	0.2539/**-01.44**%	0.2442/**-05.20**%	0.2403/**-06.72**%
ME14data				
QE-Cxt	0.3491	0.3402/**-02.55**%	0.3365/**-03.61**%	0.3354/**-03.92**%
QE-RL	0.3547	0.3529/**-00.51**%	0.3352/**-05.82**%	0.3327/**-06.20**%
QE-Cxt-RL	0.3906	0.3827/**-02.02**%	0.3769/**-03.51**%	0.3788/**-03.02**%

The approaches to index and retrieve visual features were proposed in Section 5.2. The primary difference is the use of query expansion to reconstruct the hyperlinking query. All these experiments use the strategies whose effectiveness was demonstrated in Section 6.1, including QE-RL-R, QE-Cxt and QE-Cxt-RL and the corresponding runs associated with the video-level features (META). The hyperlinking construction follows the strategies proposed in Section 4.2.

Table 6.10 show hyperlinking results using late fusion on spoken information and segment-level features. Table 6.11 describes the results using video-level features. In general, all the results are worse than those of the corresponding baselines. In Chapter 5, we concluded that the visual multimodal features can be complementary to spoken information to represent user's understanding of the video data. The current experiments demonstrate the relatively poor ability of segment-level features to represent semantically meaningful information compared with the other two methods: the expanded query and video-level features. The success of spoken information and video-level features further suggests that users make their judgement on video relevance based overall interpretation of the available information rather than specific scene objects or background.

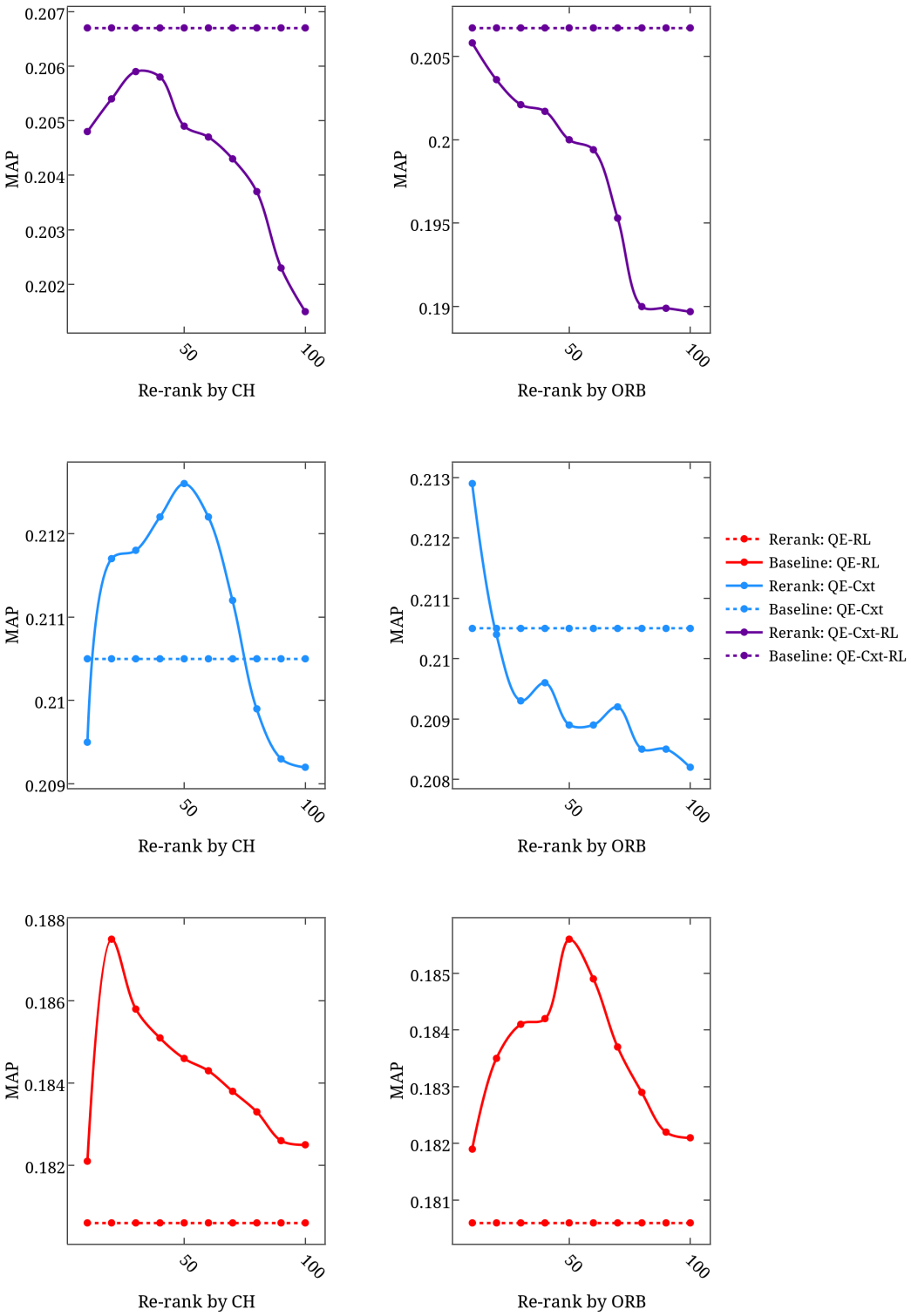


Figure 6.15: An analysis of the re-ranking strategy in terms of MAP for ME13data using QE-Cxt, QE-RL-R and QE-Cxt-RL

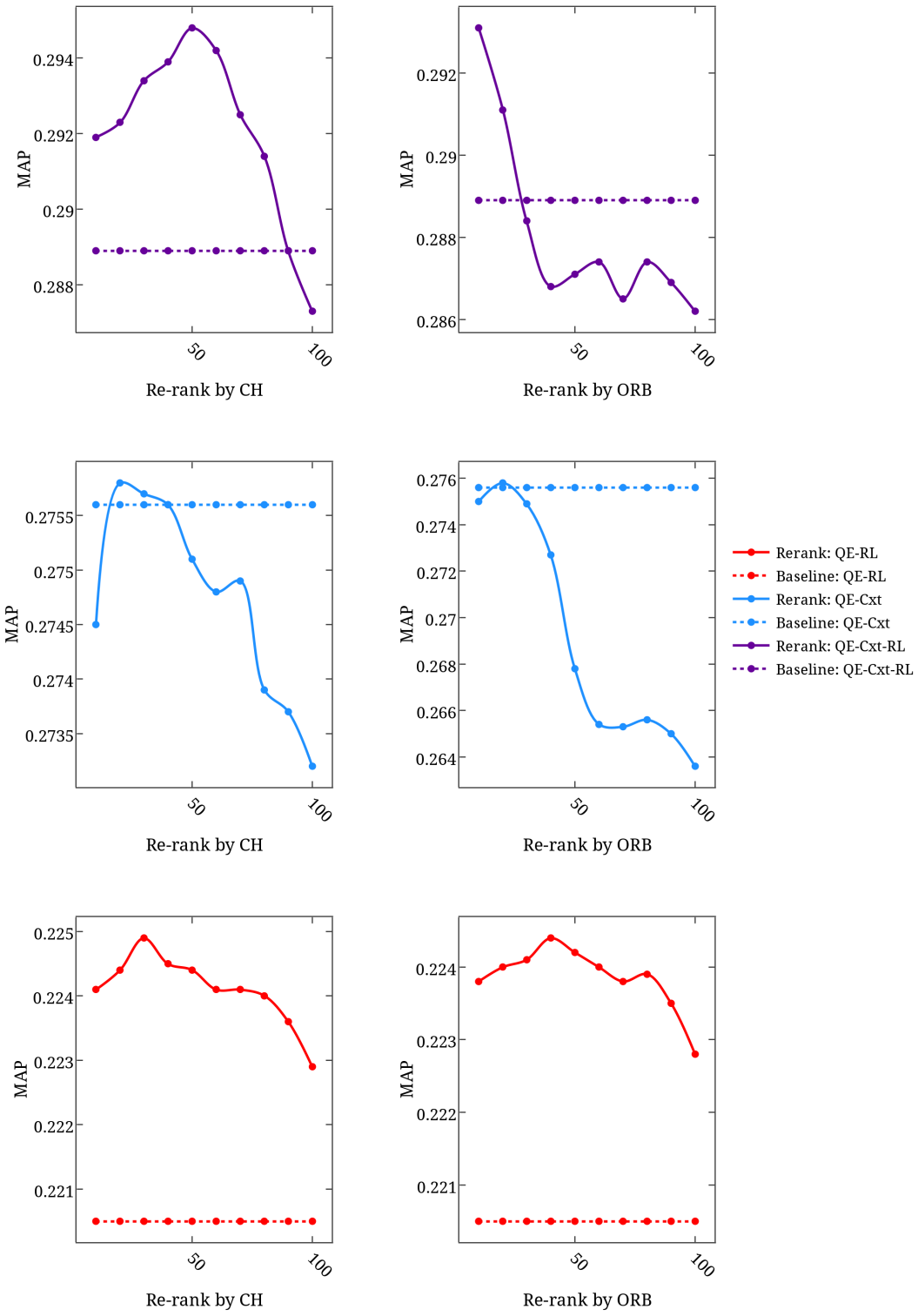


Figure 6.16: An analysis of the re-ranking strategy in terms of MAP for ME14data using QE-Cxt, QE-RL-R and QE-Cxt-RL

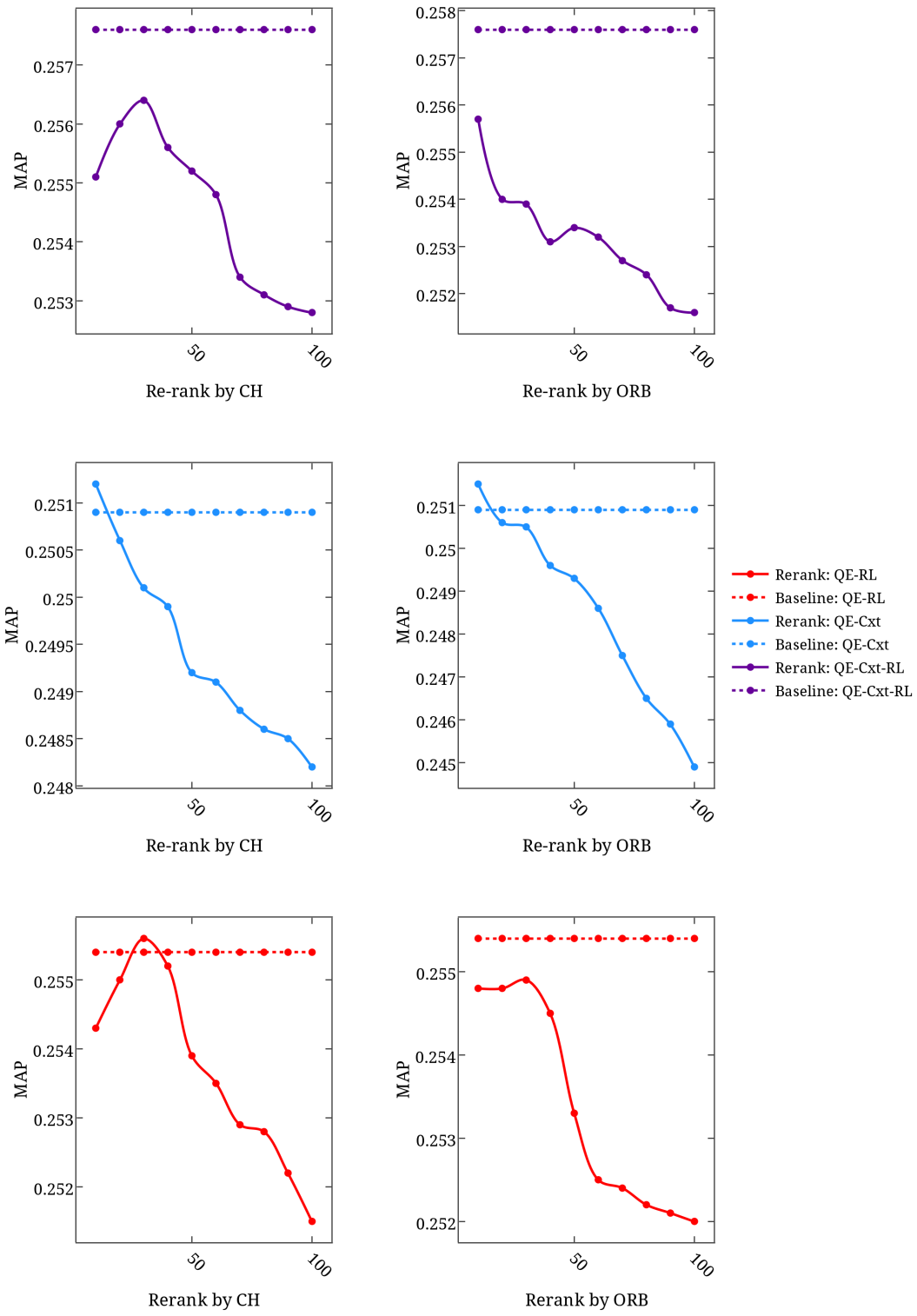


Figure 6.17: An analysis of the re-ranking strategy in terms of MAP for ME13data using H-QE-Cxt, H-QE-RL-R and H-QE-Cxt-RL

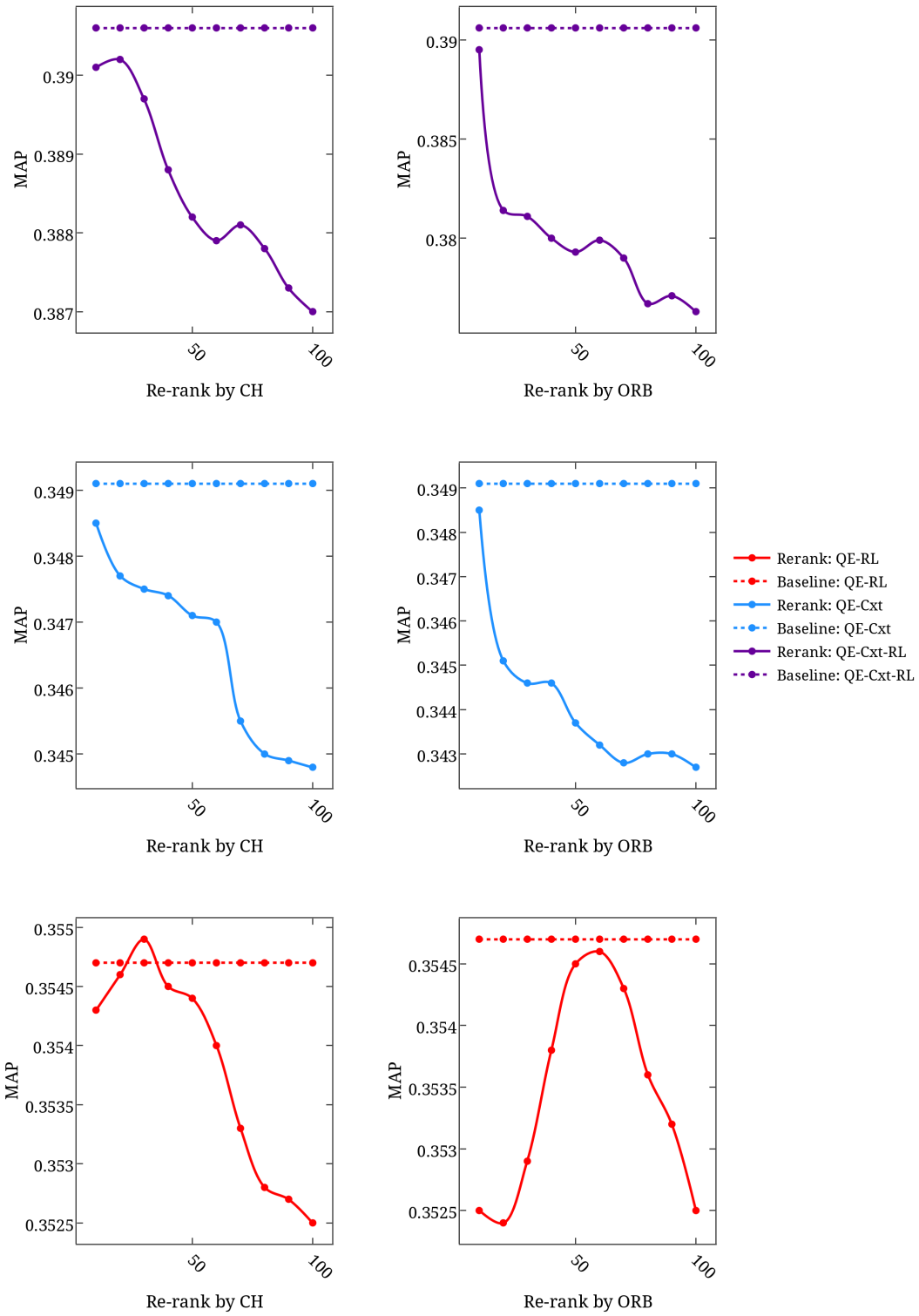


Figure 6.18: An analysis of the re-ranking strategy in terms of MAP for ME14data using H-QE-Cxt, H-QE-RL-R and H-QE-Cxt-RL

Figures 6.15 and 6.16 show the hyperlinking results in terms of MAP using the re-ranking at top R (R is from 10 to 100) retrieved segments. Figures 6.17 and 6.18 illustrate the results using the hierarchy hyperlinking model. We conclude that a combination of segment-level features and query expansion strategies can improve the hyperlinking results, as shown in Figures 6.15 and 6.16. Both figures confirm that the results of QE-RL can be improved by using CH and ORB features. However, in Figures 6.17 and 6.18, the corresponding approaches achieve only a slight improvement. For the approach QE-Cxt, we can observe a similar conclusion in Figures 6.15 and 6.16, that the re-ranking strategy can achieve better performance. However, when combined with the video-level features, the best results show only slight improvement compared with the corresponding baseline. When considering the QE-Cxt-RL, the approach of H-QE-Cxt-RL, the baseline in Figures 6.17 and 6.18, still achieves the best MAP value (0.2576 for ME13data and 0.3906 for ME14data). These results are superior to the best MAP values using the re-ranking strategy, which are 0.2059 and 0.2948 respectively in Figures 6.17 and 6.18.

6.3.4 Discussion

This section described a hyperlinking strategy based on combining visual features with query anchor analysis. We applied the approaches presented in Chapter 5 to fuse multimodal information. Both segment-level and video-level features were used, including metadata (META), high-level concept (CPT), colour histogram (CH) and ORB descriptors (ORB). The hyperlinking query was created according to the methodologies proposed in Section 6.2. The approaches included using context information to enrich the query content (QE-Cxt), and using pseudo relevant feedback and a late fusion scheme to combine hyperlinking results (QE-RL).

In Section 6.3.1, the hierarchy hyperlinking model was applied to QE-Cxt and QE-RL. The experiments revealed that with video-level features, QE-RL achieved a better MAP value in most cases. Considering the experiments in Section 6.2.2, we conclude that QE-RL can retrieve more relevant segments with relatively lower ranks. These relevant segments were located at higher ranks when being fused with video-level features. For QE-Cxt, we observe that the advantage of QE-Cxt was a better precision rate. The experimental results demonstrated that even when having relatively low MAP values, the P@5 values of QE-Cxt are still superior to QE-RL.

We showed that the QE-Cxt and QE-RL approaches are two independent directions to improve the hyperlinking query. This means that a combination of these two methods could exploit both their advantages. Therefore, in Section 6.3.2, we proposed a hyperlinking framework using both the query anchor analysis method and hierarchy hyperlinking model. This combined model uses the context information and pseudo retrieval to improve the initial retrieval, and then the video-level features (META) were applied to generate the final hyperlinking results. The experiments showed that this model achieves the best MAP and P@5 values compared with all previous results.

In Section 6.3.1, we noted that fusing META and CPT together always reduced the hyperlinking effectiveness, while using only META achieved the best results. In Section 6.3.3, we investigated different approaches to the use of segment-level features (CH and ORB), including using the late fusion scheme and re-ranking strategy. The experiments suggested that direct fusion of segment-level features reduces the hyperlinking retrieval, while re-ranking the top R results according to segment-level features can improve the results in some cases.

The experiments also showed that the best performance was obtained by the model proposed in Section 6.3.2, which involved only video-level features. The experimental investigation proposed in Section 6.3.1 and 6.3.3 supported the

conclusion that the ability of multimodal features to describe cognitive information varies. From Chapter 5, we believe that the reason for the ineffectiveness of segment-level features was the poor initial retrieval. After using an expanded query, the improved results shown in Figure 6.15 and 6.16 demonstrate that the segment-level feature is an alternative and complementary source of multimodal information. Its effectiveness, however, is lower than that of the video-level features. That revealed the benchmark of relevance judgement when users watch a pair video segments: users prefer semantically meaningful information, i.e. high-level concepts reflected by the visual descriptors, rather than the low-level visual descriptors which are often used in computer vision. Thus, the manually created metadata had the best ability to represent the corresponding high-level concepts. The automatically created concepts produced better performance than low-level visual features, but were less effective than metadata information.

6.4 Chapter Conclusion

This chapter investigated the use of query anchor analysis to improve multimedia hyperlinking. The approaches used spoken information to enrich the content of the hyperlinking query. Three methodologies were introduced:

- Use the spoken terms in the segment segments around the query anchor. (QE-Cxt)
- Use the spoken terms located at the top R results of the initial retrieval. (QE-RE)
- Use a late fusion scheme to integrate the hyperlinking results retrieved by the pseudo queries. The pseudo queries are determined by the top R results of the initial retrieval. (QE-RL)

We used a simple and efficient query expansion strategy described in [RJ94] to select the best K terms from the expanded resources. Section 6.2 illustrated the workflow of each approach.

Section 6.2.2 outlined the experimental results in terms of MAP. For QE-Cxt, the segment size P was set from 60 to 480 seconds, and the number of merged terms K was from 20 to 300. We concluded that QE-Cxt outperforms the baseline defined in Chapter 4 when applying these parameters. For QE-RE and QE-RL, R was from 5 to 40, and K was from 5 to 40. The results suggested that when $R = 5$, the algorithms can achieve the best hyperlinking performance. We introduced two methodologies to estimate the late fusion weights for QE-RL, one using equal weights (QE-RL-E) and one using rank-normalised scores (QE-RL-R). The experimental results in Table 6.6 demonstrate QE-RL-R, which applies rank-normalised scores to fuse the expanded retrieval lists, achieves better hyperlinking quality.

Section 6.3 combined the approaches to integrate multimodal features introduced in Chapter 5 with the query anchor analysis. The experiments revealed that QE-Cxt and QE-RL had their advantages respectively. The former increased the precision rate, while the latter ranked an increase number of relevant segments at lower ranks. After being fused with video-level features, QE-RL achieved a better MAP value comparing with QE-RE.

The primary contribution of this chapter is the hyperlinking framework proposed in Section 6.3.2. This integrates the query anchor analysis method with the hierarchy hyperlinking model. Experiments with this framework demonstrated better hyperlinking performance compared with all other methodologies.

This chapter also continues the topic of multimodal feature analysis introduced in Chapter 5. The experiments answer the question of which feature, high-level or low-level, is better for hyperlinking effectiveness. Both video-level features, metadata and visual concepts provided by University of Oxford, are high-level,

and both segment-level features, colour histogram and ORB descriptors using the BoVW model are low-level. We conclude that the high-level features have the advantage of representing semantically meaningful information when users watch a video segment. The high-level features always improve the hyperlinking performance, while the low-level features have limited ability to complement the multimodal information.

Our experiments leave some open issues for further research. Firstly, we did not determine the optimal parameters for QE-Cxt. Our experiments suggest that the optimal parameters are $K = 120$ and $P = 200$ for ME13data, and $K = 200$ and $P = 300$ for ME14data. In the later experiments, our conclusion is based on the optimal parameters in each collection respectively. However, the difference between parameters for the two collections implies that using the ME13data collection as the training set could provide inaccurate parameter estimation for the ME14data collection.

The success of QE-Cxt is based on the hypothesis of accepting the hyperlinks within the query video. Section 6.2.3 demonstrates that QE-Cxt can retrieve more relevant segments within the query video compared with the other strategies. This raises a concern of the validation of this kind of hyperlink. In MediaEval 2013 and 2014, the hyperlinking task accepted hyperlinks within the query video as valid. However, a hyperlink within the collection is more useful from the perspective of extending the user browsing experience. This leaves an open issue of how the approach using context information performs when hyperlinks should only link to segments in different videos.

Chapter 7

Thesis Conclusion

Multimedia hyperlinking is a research area within the field of content-based multimedia information retrieval. Since MediaEval 2012, it has grown as a special topic/task and has attracted more and more researchers. In TRECVid 2015, video hyperlinking is one of the primary tasks and attracts participants from more than 11 research groups. The experimental investigation in this thesis was concluded during the rapid development of video hyperlinking research. It can be a bridge between state-of-the-art investigations focusing on multimodal information retrieval and video hyperlinking techniques in the future.

7.1 Research Conclusion

The hyperlinking framework presented in this thesis consisted of three primary components: target-segment identification, hyperlinking construction and hyperlinking query generation. In the following, we discuss the contribution of multimodal features in each component respectively.

Target Segment Identification

We conclude that identifying potentially linked video segments is the fundamental issue of video hyperlinking, as target segments are document units in the proposed hyperlinking system. We used spoken transcripts in the previous experimental investigation to identify relevant potential target segments. We used both LIUM and LIMSI transcripts provided by the MediaEval Search and Hyperlinking task. Indexing and searching spoken information used two classic weighting models: *TF-IDF* and *BM25*.

The experimental investigation in Chapter 4 revealed that LIMSI outperformed LIUM. Our research made a further investigation into how to use LIMSI transcripts to segment video streams more efficiently. The investigation leveraged sentence identification in LIMSI transcripts. We implemented the video segmentation strategies using a fixed sliding window and dynamic sliding window associated with the lexical information. The experimental results showed that using the lexical information achieved better hyperlinking quality.

Our experimental investigation also focused on the optimal size of the sliding window (the optimal size of target segments). According to the experiments, we believe that selecting a moderate length for the sliding window is critical to improving hyperlinking performance. A short sliding window could contain insufficient multimodal features while an overlong one could contain redundant features that decrease the hyperlinking performance.

Finally, we compare the performances of the *TF-IDF* and *BM25* weighting models. Generally, *BM25* is superior to *TF-IDF* when applying a default parameter setting. In the last section of Chapter 4, a further investigation is on the optimal parameter selection for both ME13data and ME14data. We identified an appropriate range of *BM25* parameters.

Hyperlinking Construction

Multimodal features were used in the hyperlinking construction process proposed in this thesis, including not only spoken transcripts, but also low-level visual descriptors, and high-level concepts. We used color histograms and ORB descriptors to represent the visual information. The high-level concepts included the video metadata and the high-level visual concepts provided by the MediaEval workshop.

The experimental investigation in the thesis suggested that using multimodal features could improve hyperlinking performance. The effectiveness of these features, however, varies. Using the high-level concepts achieved the best hyperlinking results compared with all others. Moreover, the video metadata outperformed the high-level visual concepts. Experimental investigation showed that spoken transcripts were more effective than the low-level visual descriptors.

To integrate multimodal features, we investigated strategies to estimate the optimal fusion weights for the linear late fusion scheme. The strategies included 1) using a supervised learning solution to optimize the fusion weights on the training data collection, and applying the results to the test data collection; and 2) using an unsupervised solution referred as to MDM, whose effectiveness was demonstrated in TRECVid collection according to [Wil09]. The experimental investigation showed that it was difficult to identify an appropriate training collection for both ME13data and ME14data. Thus, using the MDM unsupervised learning solution achieved a better performance. We used the grid search strategy to identify the optimal fusion weights between the metadata and the spoken transcripts in both ME13data and ME14data. We concluded that the optimal solution was different for each collection. A primary reason is that the spoken transcripts were less representative when describing the video content in the

ME14data, and thus, hyperlinking required complementary information from the metadata.

Hyperlinking Query Generation

In Section 6.2, we concluded that the primary difference between video hyperlinking and information retrieval was how to construct a hyperlinking query. A hyperlinking system receives no user input as a query. Instead, the system has to predict what users are potentially interested in when creating hyperlinks from a query anchor. In Chapter 6, we applied various query expansion strategies to improve hyperlinking performance.

Our methodologies, in general, can be categorised as using either early fusion and late fusion to enrich the hyperlinking query. Early fusion used the context information around the query anchor and the initial retrieval results when extracting potentially relevant words to recreate the query anchor. Late fusion directly fused the initial retrieval results to determine the final ranked list. Later, we integrated multimodal features with query expansion strategies.

The experimental investigation concluded that early fusion was effective at improving the precision rate. Late fusion can retrieve more relevant documents from the ground truth pool with a lower rank. Based on this, we proposed our hyperlinking framework with multimodal features. This model applied early fusion to enrich the query anchor, used late fusion to fuse the initial retrieval to incorporate more relevant segments, and finally applied the metadata information to re-rank the hyperlinking list. The experimental investigation showed that the hyperlinking model achieved the best hyperlinking performance compared with all the other solutions proposed in this thesis.

The experimental investigation in this thesis contributes to research in this field as follows:

- We present a methodology to create the ground truth for hyperlinking evaluation using crowdsourcing. Engaging crowdsourcing workers can provide effective human annotations for a large data collection. In this thesis, we described the required workflow including publishing crowdsourcing assignment, accepting the assignment, and constructing the ground truth.

- We investigated how multimodal features performed in video hyperlinking. Our research not only focused on the effectiveness of individual feature but also the methodologies used to integrate multimodal features. The video summarisation information (the video-level features) is critical to improving hyperlinking performance. These features should be associated with other segment-level features (for example, spoken transcripts) to identify the potentially interesting points in a video accurately. The low-level visual features do not sufficiently represent relevant information in terms of human perspective.

- We presented a hyperlinking framework in Chapter 6 using multimodal information, query expansion strategies, and the integration of multimodal features. This approach outperformed all other approaches investigated in this thesis.

7.2 Future Work

This section presents some open issues for our future work, which will focus on future investigation of multimodal features.

- **Low-level Visual Feature** The low-level visual features applied in this thesis since they are widely investigated in multimedia retrieval. However, research in multimedia information retrieval presented a large number of low-level visual features which could benefit multimedia hyperlinking.

Thus, it would be interesting to investigate other low-level visual features to interpret the visual content in a different way.

- **High-level Visual Feature** High-level visual features are video-level in this thesis. Although the experimental investigation showed its effectiveness, we believe that segment-level high-level features could further benefit the multimedia hyperlinking. Existing research in [SSH14b] described the methodology of extracting segment-level concepts from spoken transcripts and showed good performance in ME13data. Thus, our further investigation will focus on extracting visual concepts from each target segment to describe the corresponding content.

- **Data Fusion** Multimodal feature analysis uses data fusion to integrate the hyperlinking results retrieved from different modalities. This thesis has focused on linear late fusion. Our future work will carry out further investigation into determining both linear fusion weights and non-linear data fusion strategies.

- **Parameter Selection** The experimental investigation of this thesis has revealed that parameter optimization is a primary issue in video hyperlinking, especially for query anchor expansion. Although we suggested an appropriate range of various parameters for ME13data and ME14data, it would be essential to investigate how these generalise to other data collections.

Appendix A

An Analysis of ME13data and ME14data Ground Truth

Table A.1: An overview of ME13data ground truth in terms of Mturk users' perspective. (Pos.: Both users regard the ground truth as relevant. Neg.: Both users regard the ground truth as irrelevant. Un.: Only one user regards the ground truth as relevant, while the other regard it as irrelevant)

Query ID	**Pos.**	**Neg.**	**Un.**	**Samples**
Q1	110	237	15	362
Q2	98	171	15	284
Q3	109	214	23	346
Q4	128	205	13	346
Q5	60	234	24	318
Q6	76	241	18	335
Q7	105	230	7	342
Q8	159	147	19	325
Q9	80	218	16	314
Q10	91	236	14	341
Q11	44	266	17	327
Q12	107	245	15	367
Q13	118	211	22	351
Q14	75	237	18	330
Q15	72	253	24	349
Q16	90	222	20	332
Q17	34	290	6	330
Q18	37	309	7	353
Q19	22	292	19	333
Q20	22	300	10	332
Q21	166	209	15	390
Q22	209	95	21	325
Q23	46	279	8	333
Q24	123	196	14	333
Q25	62	224	12	298
Q26	87	196	18	301
Q27	93	193	16	302
Q28	105	200	16	321
Q29	96	251	6	353
Q30	83	211	6	300

Table A.2: An overview of ME14data ground truth in terms of Mturk users' perspective. (Pos.: Both users regard the ground truth as relevant. Neg.: Both users regard the ground truth as irrelevant. Un.: Only one user regards the ground truth as relevant, while the other regard it as irrelevant)

Query ID	**Pos.**	**Neg.**	**Un.**	**Samples**
Q1	52	396	16	464
Q2	60	405	9	474
Q3	43	386	2	431
Q4	58	354	5	417
Q5	36	385	16	437
Q6	108	308	6	422
Q7	103	310	3	416
Q8	40	223	4	267
Q9	84	360	11	455
Q10	46	336	26	408
Q11	82	273	23	378
Q12	63	357	8	428
Q13	83	294	15	392
Q14	36	375	7	418
Q15	80	298	12	390
Q16	63	374	26	463
Q17	72	264	19	355
Q18	47	358	13	418
Q19	17	411	14	442
Q20	95	354	8	457
Q21	66	382	19	467
Q22	81	339	30	450
Q23	122	262	17	401
Q24	31	355	19	405
Q25	27	359	3	389
Q26	39	328	30	397
Q27	107	265	3	375
Q28	26	371	22	419
Q29	47	389	15	451
Q30	74	170	10	254

Appendix B

Comparsion of Proposed Hyperlinking Solution with Other Investigations

Figure 6.11 concluded our final hyperlinking design in this thesis with the best MAP 0.2576 in ME13data and 0.3906 in ME14data. In this part, we compare these results with other hyperlinking runs reviewed in Sections 2.4.3 and 2.4.3. Furthermore, TRECVid 2015 proposed the Video Hyperlinking task (LNK) using ME14data[1]. We also submitted our experimental runs using the hyperlinking system illustrated in Figure 6.11. Thus, we also compare our results with those from other participants. All the tables use the term "PROPOSED" to represent our results using the hyperlinking system illustrated in Figure 6.11.

[1]TRECVid 2015 and MediaEval 2014 used the same collection in hyperlinking task with different query set.

Table B.1: A comparison between our solution and the results from all other participants in MediaEval 2013

TEAM	MAP
Idiap2013 [BPHPB13]	0.5172
PROPOSED	**0.2576**
DCU [CJO13]	0.2354
LinkedTV13 [SHC+13]	0.2321
TOSCA-MP2013 [LSB13]	0.1887
UTwente [SAO13]	0.0609
soton-wais2013[PHS+13]	0.0594
HITSIRISA [GSGS13]	0.0474
MMLab [NNMdW13]	0.0376
UPC [VTAN13]	0.0240

Table B.2: A comparison between our solution and the results from all other participants in MediaEval 2013

TEAM	MAP
CUNI [GPKL14]	4.1824
PROPOSED	**0.3906**
LINKEDTV2014 [PMS+14]	0.2524
DCU [CJO14]	0.0791
JRS [BS14]	0.0556
IRISAKUL [SGSM14]	0.0335
DCLab [PFS14]	0.0135

Table B.3: A comparison between our submission and the results from all other participants in TRECVid 2015 hyperlinking task. (MAiSP [RJ15] is a new evaluation metric used in TRECVid 2015 hyperlinking task.)

TEAM	MAP	TEAM	MAiSP
CMU	0.4623	**PROPOSED**	**0.2718**
PROPOSED	**0.3044**	CMU	0.2690
EURECOM	0.2179	EURECOM	0.2020
VIREO	0.1890	VIREO	0.1782
CUNI	0.1441	CUNI	0.1311
ORAND	0.1071	IRISA	0.0792
IRISA	0.0873	ORAND	0.0563
iip	0.0419	iip	0.0338
Metu	0.0294	Metu	0.0297
TUZ	0.0177	TUZ	0.0238

Bibliography

[ACD+98] James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. "Topic Detection and Tracking Pilot Study Final Report". In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, February 1998.

[AEOJ13] Robin Aly, Maria Eskevich, Roeland Ordelman, and Gareth J.F. Jones. "Adapting Binary Information Retrieval Evaluation Metrics for Segment-based Retrieval Tasks". *Computing Research Repository*, December 2013.

[AHESK10] Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. "Multimodal Fusion for Multimedia Analysis: a Survey". *Multimedia Systems*, 16:345–379, 2010.

[AKRR99] Dave Abberley, David Kirby, Steve Renals, and Tony Robinson. "The THISL Broadcast News Retrieval System". In *Proceedings of ESCA on Accessing Information in Spoken Audio*, pages 19–24, Cambridge, UK, April 1999.

[AMC+12] Robin Aly, Kevin McGuinness, Shu Chen, Noel E. O'Connor, Ken Chatfield, Omkar Parkhi, Relja Arandjelovic, Andrew Zisserman, Basura Fernando, and Tinne Tuytelaars. "AXES at TRECVid 2012: KIS, INS, and MED". In *Proceedings of TREC Video Retrieval Evaluation*, Maryland, USA, November 2012.

[ASD12] Tim Althoff, Hyun Oh Song, and Trevor Darrell. "Detection Bank: an Object Detection based Video Representation for Multimedia Event Recognition". In *Proceedings of the 20th ACM International Conference on Multimedia*, pages 1065–1068, 2012.

[BBL+08] Hervé Bredin, Daragh Byrne, Hyowon Lee, Noel E. O'Connor, and Gareth J.F. Jones. "Dublin City University at the TRECVid 2008 BBC Rushes Summarisation Task". In *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*, pages 45–49, Vancouver, Canada, 2008.

[BD90] Hans Peter Brondmo and Glorianna Davenport. "Creating and viewing the Elastic Charles-a hypermedia journal". *Hypertext, State of the Art*, 1990.

[BG98] Suresh Balakrishnama and Aravind Ganapathiraju. "Linear Discriminant Analysis - a Brief Tutorial". *International Symposium on Information Processing*, 1998.

[BHdR11] Marc Bron, Bouke Huurnink, and Maarten de Rijke. "Linking Archives using Document Enrichment and Term Selection". In *Proceedings of the 15th International Conference on Theory and Practice of Digital Libraries: Research and Advanced Technology for Digital Libraries (TPDL '11)*, pages 360–371, Berlin, German, September 2011.

[BMI12] Andrzej Białecki, Robert Muir, and Grant Ingersoll. "Apache Lucene 4". In *SIGIR 2012 Workshop on Open Source Information Retrieval*, pages 17–24, Portland, USA, 2012.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". *Journal of Machine Learning Research*, 3:993–1022, 2003.

[BPHPB13] Chidansh A. Bhatt, Nikolaos Pappas, Maryam Habibi, and Andrei Popescu-Belis. "IDIAP at Mediaeval 2013: Search and Hyperlinking Task". In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain, October 2013.

[BS14] Werner Bailer and Harald Stiegler. "JRS at Search and Hyperlinking of Television Content Task". In *Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain, October 2014.

[BTVG06] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. "SURF: Speeded up Robust Features". In *European Conference on Computer Vision*, pages 404–417. 2006.

[Bus45] Vannevar Bush. "As We May Think". In *The Atlantic*, 1945.

[BUS10] Christian Beecks, Merih Seran Uysal, and Thomas Seidl. "Signature Quadratic Form Distance". In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 438–445, 2010.

[BYRN99] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *"Modern Information Retrieval"*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[CdCC+05] B Cardoso, Fausto de Carvalho, Luis Carvalho, Gabriel Fernàndez, Paulo Gouveia, Benoit Huet, Joakim Jiten, Alejandro López, Bernard Mérialdo, Antonio Navarro, et al. "Hyperlinked video

with moving objects in digital television". In *IEEE International Conference on Multimedia and Expo*, pages 486–489, 2005.

[CEJO14] Shu Chen, Maria Eskevich, Gareth J.F. Jones, and Noel E. O'Connor. "An Investigation into Feature Effectiveness for Multimedia Hyperlinking". In *MultiMedia Modeling*, volume 8326, pages 251–262. 2014.

[CH05] Michael G. Christel and Alexander G. Hauptmann. "The Use and Utility of High-level Semantic Features in Video Retrieval". In *Proceedings of the 4th International Conference on Image and Video Retrieval*, pages 134–144, Berlin, Heidelberg, 2005.

[CJO12] Shu Chen, Gareth J.F. Jones, and Noel E. O'connor. "DCU Linking Runs at Mediaeval 2012: Search and Hyperlinking Task". In *Proceedings of the MediaEval 2012 Multimedia Benchmark Workshop*, Pisa, Italy, October 2012.

[CJO13] Shu Chen, Gareth J.F. Jones, and Noel E. O'connor. "DCU Linking Runs at Mediaeval 2013: Search and Hyperlinking Task". In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain, October 2013.

[CJO14] Shu Chen, Gareth J.F. Jones, and Noel E. O'connor. "DCU Linking Runs at Mediaeval 2014: Search and Hyperlinking Task". In *Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain, October 2014.

[CJSW01] Heng-Da Cheng, XH Jiang, Ying Sun, and Jingli Wang. "Color Image Segmentation: Advances and Prospects". *Pattern recognition*, 34:2259–2281, 2001.

[CLSF10] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. "BRIEF: Binary Robust Independent Elementary Features". In *European Conference on Computer Vision*, pages 778–792. 2010.

[CLVZ11] Ken Chatfield, Victor Lempitsky, Andrea Vedaldi, and Andrew Zisserman. "The Devil is in the Details: An Evaluation of Recent Feature Encoding Methods". In *Proceedings of the British Machine Vision Conference*, volume 76, pages 1–12, Dundee, UK, September 2011.

[CMA+14] Shu Chen, Kevin McGuinness, Robin Aly, Noel E. OConnor, and Tinne Tuytelaars. "AXES @ TRECVid 2014: Instance Search". In *Proceedings of TREC Video Retrieval Evaluation*, 2014.

[CRM03] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. "Kernel-based Object Tracking". volume 25, pages 564–577, 2003.

[CSVZ14] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Return of the Devil in the Details: Delving Deep into Convolutional Nets". *Computing Research Repository*, 1405.3531, 2014.

[CZ13] Ken Chatfield and Andrew Zisserman. "VISOR: Towards On-the-fly Large-scale Object Category Retrieval". In *Proceedings of the 11th Asian Conference on Computer Vision - Volume Part II*, pages 432–446, Daejeon, Korea, 2013.

[DACBJ99] Jonathan Dakss, Stefan Agamanolis, Edmond Chalom, and V Michael Bove Jr. "Hyperlinked Video". In *Photonics East (ISAM, VVDC, IEMB)*, pages 2–10. International Society for Optics and Photonics, 1999.

[Dak99] Jonathan Dakss. *"HyperActive: an automated tool for creating hyperlinked video"*. PhD thesis, Massachusetts Institute of Technology, 1999.

[DH72] Richard O. Duda and Peter E. Hart. "Use of the Hough Transformation to Detect Lines and Curves in Pictures". *Communications of the ACM*, 15:11–15, January 1972.

[DHS73] Richard O. Duda, Peter E. Hart, and David G. Stock. *"Pattern Classification and Scene analysis"*. 1973.

[DJLW08] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. "Image Retrieval: Ideas, Influences, and Trends of the New Age". *ACM Computing Surveys (CSUR)*, 40:5:1–5:60, 2008.

[DNDVD+12] Tom De Nies, Pedro Debevere, Davy Van Deursen, Wesley De Neve, Erik Mannens, and Rik Van de Walle. "Ghent University-IBBT at MediaEval 2012 Search and Hyperlinking: Semantic Similarity using Named Entities". In *Proceedings of the MediaEval 2012 Multimedia Benchmark Workshop*, Pisa, Italy, October 2012.

[EAL12] Maria Eskevich, Robin Aly, and Martha Larson. "The Search and Hyperlinking Task at Mediaeval 2012". In *Proceedings of the MediaEval 2012 Multimedia Benchmark Workshop*, Pisa, Italy, October 2012.

[EAO+14] Maria Eskevich, Robin Aly, Roeland Ordelman, David N. Racca, Shu Chen, and Gareth J.F. Jones. "The Search and Hyperlinking Task at Mediaeval 2014". In *Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain, 2014.

[EJC+13] Maria Eskevich, Gareth J.F. Jones, Shu Chen, Robin Aly, and Roeland Ordelman. "Search and Hyperlinking Task at MediaEval 2013". In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain, October 2013.

[EOS12] Sandra Garcia Esparza, Michael P OMahony, and Barry Smyth. "Mining the Real-time Web: a Novel Approach to Product Recommendation". *Knowledge-Based Systems*, 29:3–11, May 2012.

[Fau93] Olivier Faugeras. *"Three-dimensional Computer Vision: a Geometric Viewpoint"*. MIT Press, Cambridge, MA, USA, 1993.

[FCH+08] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. "LIBLINEAR: a Library for Large Linear Classification". *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[FFP05] Li Fei-Fei and Pietro Perona. "A Bayesian Hierarchical Model for Learning Natural Scene Categories". In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 524–531, 2005.

[FHHD92] Andrew M. Fountain, Wendy Hall, Ian Heath, and Hugh C. Davis. *"Hypertext: Concepts, Systems and Applications"*. 1992.

[Fis36] Ronald A Fisher. "The Use of Multiple Measurements in Taxonomic Problems". *Annals of eugenics*, 7:179–188, 1936.

[FS94] Edward A. Fox and Joseph A. Shaw. "Combination of Multiple Searches". *NIST Special Publication*, pages 243–243, 1994.

[FV07] Mohamed Farah and Daniel Vanderpooten. "An Outranking Approach for Rank Aggregation in Information Retrieval". In *Proceed-*

ings of the 30th Annual International ACM SIGIR Conference, pages 591–598, 2007.

[GGS12] Camille Guinaudeau, Guillaume Gravier, and Pascale Sebillot. "IRISA at MediaEval 2012: Search and Hyperlinking Task". In *Proceedings of the MediaEval 2012 Multimedia Benchmark Workshop*, Pisa, Italy, October 2012.

[GLJ11] Debasis Ganguly, Johannes Leveling, and Gareth J.F. Jones. *"Query Expansion for Language Modeling using Sentence Similarities"*. Springer, 2011.

[Goo00] Abby Goodrum. "Image Information Retrieval: An Overview of Current Research". *Informing Science*, 3:63–66, 2000.

[GP13] Petra Galuscakova and Pavel Pecina. "CUNI at MediaEval 2014 Search and Hyperlinking Task: Search Task Experiments". In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain, October 2013.

[GP14] Petra Galuscakova and Pavel Pecina. "Experiments with Segmentation Strategies for Passage Retrieval in Audio-visual Documents". In *Proceedings of International Conference on Multimedia Retrieval (ICMR '14)*, Glasgow, Scotland, UK, April 2014.

[GPKL14] Petra Galuscakova, Pavel Pecina, Martin Krulis, and Jakub Lokoc. "CUNI at Mediaeval 2014 Search and Hyperlinking Task: Visual and Prosodic Features in Hyperlinking". In *Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain, October 2014.

[GSGS13] Camille Guinaudeau, Anca-Roxana Simon, Guillaume Gravier, and Pascale Sebillot. "HITS and IRISA at Mediaeval 2013: Search and Hyperlinking Task". In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain, October 2013.

[HS88] Chris Harris and Mike Stephens. "A Combined Corner and Edge Detector". In *Proceedings of 4th Alvey Vision Conference*, volume 15, pages 147–151, Manchester, UK, 1988.

[HSE+95] James Hafner, Harpreet S. Sawhney, William Equitz, Myron Flickner, and Wayne Niblack. "Efficient Color Histogram Indexing for Quadratic Form Distance Functions". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:729–736, 1995.

[HYL07] Alexander Hauptmann, Rong Yan, and Wei-Hao Lin. "How Many High-level Concepts will Fill the Semantic Gap in News Video Retrieval?". In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval (CIVR '07)*, pages 627–634, Amsterdam, Netherlands, 2007.

[Ing96] Peter Ingwersen. "Cognitive Perspectives of Information Retrieval Interaction: Elements of a Cognitive IR Theory". *Journal of Documentation*, 52:3–50, 1996.

[INN03] Giridharan Iyengar, Harriet J Nock, and Chalapathy Neti. "Audio-Visual Synchrony for Detection of Monologues in Video Archives". In *Proceedings of International Conference on Multimedia and Expo*, volume 1, pages 329–332, 2003.

[JFY09] Thorsten Joachims, Thomas Finley, and Chun-NamJohn Yu. "Cutting-plane Training of Structural SVMs". *Machine Learning*, 77:27–59, 2009.

[Jon13] Gareth J.F. Jones. "An Introduction to Crowdsourcing for Language and Multimedia Technology Research". In *Information Retrieval Meets Information Visualization*, pages 132–154. 2013.

[JS13] Hideo Joho and Tetsuya Sakai. "Overview of NTCIR-10". In *Proceedings of the 10th NTCIR Conference*, NII, Tokyo, Japan, June 2013.

[JZCL08] Wei Jiang, E. Zavesky, Shih-Fu Chang, and A. Loui. "Cross-domain Learning Methods for High-level Visual Concept Classification". In *15th IEEE International Conference on Image Processing (ICIP '08)*, pages 161–164, October 2008.

[KB13] Simardeep Kaur and Dr Vijay Kumar Banga. "Content-based Image Retrieval: Survey and Comparison between RGB and HSV Model". *International Journal of Engineering Trends and Technology*, 4:575–579, 2013.

[KG10] Roman Kern and Michael Granitzer. "German Encyclopedia Alignment based on Information Retrieval Techniques". In *Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries (ECDL '10)*, pages 315–326, Glasgow, UK, September 2010.

[KS04] Yan Ke and Rahul Sukthankar. "PCA-SIFT: a More Distinctive Representation for Local Image Descriptors". In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 506–513, June 2004.

[Le13] Quoc V. Le. "Building High-level Features using Large Scale Unsupervised Learning". In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8595–8598, 2013.

[LG08] Lori Lamel and Jean-Luc Gauvain. "Speech Processing for Audio Indexing". In *Advances in Natural Language Processing*, pages 4–15. 2008.

[LHR99] B.J. Lei, Emile A. Hendriks, and M.J.T. Reinders. *"On Feature Extraction from Images"*. Information and Communication Theory Group, TUDelft, 1999.

[Lin12] Chih-long Lin. "Content-based Video Retrieval with Multi Features". pages 1248–1257, 2012.

[LLYK06] Ying Li, Shih-Hung Lee, Chia-Hung Yeh, and C-CJ Kuo. "Techniques for Movie Content Analysis and Skimming: Tutorial and Overview on Video Abstraction Techniques". *IEEE Signal Processing Magazine*, 23:79–89, 2006.

[LM01] Thomas Leung and Jitendra Malik. "Representing and Recognizing the Visual Appearance of Materials using Three-dimensional Textons". *International Journal of Computer Vision*, 43:29–44, 2001.

[Low04] David G Lowe. "Distinctive Image Features from Scale-invariant Keypoints". *International Journal of Computer Vision*, 60:91–110, 2004.

[LS13] Manisha Lumb and Poonam Sethi. "Texture Feature Extraction of RGB, HSV, YIQ and Dithered Images using GLCM, Wavelet Decomposition Techniques". *International Journal of Computer Applications*, 68:25–31, 2013.

[LSB13] Michal Lokaj, Harald Stiegler, and Werner Bailer. "TOSCA-MP at Search and Hyperlinking of Television Content Task". In *Pro-*

ceedings of the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, October 2013.

[LSFFX10] Li-Jia Li, Hao Su, Li Fei-Fei, and Eric P. Xing. "Object Bank: a High-level Image Representation for Scene Classification & Semantic Feature Sparsification". In *Advances in Neural Information Processing Systems*, pages 1378–1386, 2010.

[LSLFF14] Li-Jia Li, Hao Su, Yongwhan Lim, and Li Fei-Fei. "Object Bank: an Object-level Image Representation for High-level Visual Recognition". *International Journal of Computer Vision*, 107:20–39, 2014.

[LSP06] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories". In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2169–2178, 2006.

[LZ11] Yuanhua Lv and ChengXiang Zhai. "Lower-bounding Term Frequency Normalization". In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11)*, pages 7–16, 2011.

[Ma09] Ji-quan Ma. "Content-based Image Retrieval with HSV Color Space and Texture Features". In *Web Information Systems and Mining, 2009. WISM 2009. International Conference on*, pages 61–63, 2009.

[MBL+10] Christophe Moulin, Cécile Barat, Cédric Lemaître, Mathias Géry, Christophe Ducottet, and Christine Largeron. "Combining Text/Image in WikipediaMM Task 2009". In *Multilingual Infor-*

mation Access Evaluation II. Multimedia Experiments, pages 164–171. 2010.

[MC07] Rada Mihalcea and Andras Csomai. "Wikify!: Linking Documents to Encyclopedic Knowledge". In *Proceedings of the 16th ACM conference on Information and Knowledge Management (CIKM '07)*, pages 233–242, Lisbon, Portugal, November 2007.

[ML09] Marius Muja and David G. Lowe. "Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration". In *International Conference on Computer Vision Theory and Application*, volume 2, pages 331–340, Lisboa, Portugal, February 2009.

[MLD+06] A. Massoudi, F. Lefebvre, C. Demarty, L. Oisel, and B. Chupeau. "A Video Fingerprint Based on Visual Digest and Local Fingerprints". In *IEEE International Conference on Image Processing*, pages 2297–2300, Octobor 2006.

[MLD+14] Christophe Moulin, Christine Largeron, Christophe Ducottet, Mathias Géry, and Cécile Barat. "Fisher Linear Discriminant Analysis for Text-image Combination in Multimedia Information Retrieval". *Pattern Recognition*, 47:260–269, 2014.

[Mor81] Hans P. Moravec. "Rover Visual Obstacle Avoidance". In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, pages 785–790, San Francisco, USA, 1981.

[MRS08] C.D. Manning, P. Raghavan, and H. Schutze. *"Introduction to Information Retrieval"*. Cambridge University Press, 2008.

[MW08] David Milne and Ian H. Witten. "Learning to Link with Wikipedia". In *Proceedings of the 17th ACM conference on Information*

and Knowledge Management (CIKM '08), pages 509–518, Napa Valley, California, USA, November 2008.

[NAO12] Danish Nadeem, Robin Aly, and Roeland Ordelman. "UTwente does Brave New Tasks for Mediaeval 2012: Searching and Hyperlinking". In *Proceedings of the MediaEval 2012 Multimedia Benchmark Workshop*, Pisa, Italy, October 2012.

[NB80] Ramakant Nevatia and K. Ramesh Babu. "Linear Feature Extraction and Description". *Computer Graphics and Image Processing*, 13:257–269, 1980.

[NNMdW13] Tom D. Nies, Wesley D. Neve, Erik Mannens, and Rik V. de Walle. "Ghent University-iMinds at Mediaeval 2013: An Unsupervised Named Entity-based Similarity Measure for Search and Hyperlinking". In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain, October 2013.

[NPG13] Romain Negrel, David Picard, and Philippe-Henri Gosselin. "Web-scale Image Retrieval using Compact Tensor Aggregation of Visual Descriptors". *IEEE MultiMedia*, 20:24–33, 2013.

[OEA+15] Roeland Ordelman, Maria Eskevich, Robin Aly, Benoit Huet, and Gareth J.F. Jones. "Defining and Evaluating Video Hyperlinking for Navigating Multimedia Archives". In *Proceedings of the 24th International Conference on World Wide Web Companion*, pages 727–732, 2015.

[Oga86] Hideo Ogawa. "Labeled Point Pattern Matching by Delaunay Triangulation and Maximal Cliques". *Pattern Recognition*, 19:35–40, January 1986.

[ORMA01] T. Ojala, M. Rautiainen, E. Matinmikko, and M. Aittola. "Semantic Image Retrieval with HSV Correlograms". In *Proceedings of the Scandinavian Conference on Image Analysis*, pages 621–627, 2001.

[PCI+07] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. "Object Retrieval with Large Vocabularies and Fast Spatial Matching". In *Computer Vision and Pattern Recognition*, pages 1–8, Minneapolis, MN, USA, March 2007.

[Pet00] Milan Petkovic. *"Content-based Video Retrieval"*. University of Konstanz, 2000.

[PFS14] Zsombor Paroczi, Balint Fodor, and Gabor Szucs. "DCLab at Mediaeval2014 Search and Hyperlinking Task". In *Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain, October 2014.

[PHS+13] John Preston, Jonathon Hare, Sina Samangooei, Jamie Davies, Neha Jain, David Dupplaw, and Paul H Lewis. "A Unified, Modular and Multimodal Approach to Search and Hyperlinking Video". In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain, October 2013.

[PKPM09] George Papandreou, Athanassios Katsamanis, Vassilis Pitsikalis, and Petros Maragos. "Adaptive Multimodal Fusion by Uncertainty Compensation with Application to Audiovisual Speech Recognition". *Transactions on Audio Speech and Language Processing*, 17:423–435, Mar 2009.

[PMS+14] A. Pournaras, V. Mezaris, D. Stein, S. Eickeler, and M. Stadtschnitzer. "LinkedTV at Mediaeval 2014 Search

and Hyperlinking Task". In *Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain, October 2014.

[Por80] Martin F Porter. "An Algorithm for Suffix Stripping". *Program*, 14:130–137, 1980.

[PW10] Ofir Pele and Michael Werman. "The Quadratic-chi Histogram Distance Family". In *European Conference on Computer Vision*, pages 749–762. 2010.

[RBD+11] A. Rousseau, F. Bougares, P. Delssglise, H. Schwenk, and Y. Estssv. "LIUM's Systems for the IWSLT 2011 Speech Translation Tasks". In *Proceedings of International Workshop on Spoken Language Translation (IWSLT 2011)*, San Francisco, USA, September 2011.

[RD06] Edward Rosten and Tom Drummond. "Machine Learning for High-speed Corner Detection". In *European Conference on Computer Vision*, pages 430–443. 2006.

[RJ88] Stephen Robertson and Karen E. Sparck Jones. Document Retrieval Systems. In *"Relevance Weighting of Search Terms"*, pages 143–160, 1988.

[RJ94] Stephen E Robertson and Karen Sparck Jones. *"Simple, Proven Approaches to Text Retrieval"*. Computer Laboratory, University of Cambridge, 1994.

[RJ15] D. N. Racca and G. J. F. Jones. "Evaluating Search and Hyperlinking: An Example of the Design, Test, Refine Cycle for Metric Development". In *Proceedings of the MediaEval 2015 Workshop*, Wurzen, Germany, 2015.

[Rob90] Stephen E. Robertson. "On Term Selection for Query Expansion". *Journal of Documentation*, 46:359–364, 1990.

[RRKB11] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. "ORB: An Efficient Alternative to SIFT or SURF". In *IEEE International Conference on Computer Vision (ICCV)*, pages 2564–2571, 2011.

[RvRP81] Stephen E. Robertson, C. J. van Rijsbergen, and M. F. Porter. "Probabilistic Models of Indexing and Searching. In *Proceedings of the 3rd Annual ACM Conference on Research and Development in Information Retrieval*, pages 35–56, Cambridge, England, 1981.

[RWJ+95] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. "Okapi at TREC-3". *NIST SPECIAL PUBLICATION SP*, pages 109–109, 1995.

[SAO13] Kim Schouten, Robin Aly, and Roeland Ordelman. "Searching and Hyperlinking using Word Importance Segment Boundaries in Mediaeval 2013". In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain, October 2013.

[SC12] Sreemanananth Sadanand and Jason J. Corso. "Action Bank: a High-level Representation of Activity in Video". In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1234–1241, 2012.

[SF10] Yi Shen and Jianping Fan. "Leveraging Loosely-tagged Images and Inter-object Correlations for Tag Recommendation". In *Proceedings of the International Conference on Multimedia (MM '10)*, pages 5–14, Firenze, Italy, 2010.

[SGF+11] David Scott, Jinlin Guo, Colum Foley, Frank Hopfgartner, Cathal Gurrin, and Alan F. Smeaton. "TRECVid 2011 Experiments at Dublin City University". In *Proceedings of TREC Video Retrieval Evaluation*, 2011.

[SGSM14] Anca-Roxana Simon, Guillaume Gravier, Pascale Sebillot, and Marie-Francine Moens. "IRISA and KUL at Mediaeval 2014: Search and Hyperlinking Task". In *Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain, October 2014.

[SHC+13] Mathilde Sahuguet, Benoit Huet, Barbora Cervenkova, Evlampios Apostolidis, Vasileios Mezaris, Daniel Stein, Stefan Eickeler, JL Redondo Garcia, Raphael Troncy, and Lukas Pikora. "LinkedTV at Mediaeval 2013 Search and Hyperlinking Task". In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain, October 2013.

[SHP12] M Singha, K Hemachandran, and A Paul. "Content-based Image Retrieval using the Combination of the Fast Wavelet Transformation and the Colour Histogram". *Image Processing*, 6:1221–1226, 2012.

[SJ11] Tetsuya Sakai and Hideo Joho. "Overview of NTCIR-9". In *Proceedings of NTCIR-9 Workshop Meeting*, pages 1–7, Tokyo, Japan, December 2011.

[Sme07] Alan F. Smeaton. "Techniques Used and Open Challenges to the Analysis, Indexing and Retrieval of Digital Video". *Information Systems*, 32:545–559, 2007.

[Smi78] Alvy Ray Smith. "Color Gamut Transform Pairs". In *ACM Siggraph Computer Graphics*, pages 12–19, 1978.

[SOK09] Alan F. Smeaton, Paul Over, and Wessel Kraaij. "High-level Feature Detection from Video in TRECVid: a 5-year Retrospective of Achievements". In *Multimedia Content Analysis*, pages 1–24, 2009.

[Son09] Amol Sonawane. "Using Apache Lucene to Search Text - Easily Build Search and Index Capabilities into your Applications", August 2009. `http://www.ibm.com/developerworks/library/os-apache-lucenesearch/`.

[SQP02a] Shamik Sural, Gang Qian, and Sakti Pramanik. "A Histogram with Perceptually Smooth Color Transition for Image Retrieval". In *4th International Conferrence on Computer Vision, Pattern Recognition and Image Processing*, pages 664–667, 2002.

[SQP02b] Shamik Sural, Gang Qian, and Sakti Pramanik. "Segmentation and Histogram Generation using the HSV Color Space for Image Retrieval". In *Proceedings of International Conference on Image Processing*, volume 2, pages 589–592, 2002.

[SSH14a] Bahjat Safadi, Mathilde Sahuguet, and Benoit Huet. "Linking Text and Visual Concepts Semantically for Cross Modal Multimedia Search". In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 143–152, Barcelona, Spain, 2014.

[SSH14b] Bahjat Safadi, Mathilde Sahuguet, and Benoit Huet. "When Textual and Visual Information Join Forces for Multimedia Retrieval". In *Proceedings of International Conference on Multimedia Retrieval*, pages 265–272, Glasgow, United Kingdom, 2014.

[SSZ12] Chengyao Shen, Mingli Song, and Qi Zhao. "Learning High-level Concepts by Training a Deep Network on Eye Fixations". In *Deep*

Learning and Unsupervised Feature Learning NIPS Workshop, Lake Tahoe, USA, December 2012.

[Stu94] R. Stuckless. "Developments in real-time speech-to-text communication for people with impaired hearing". *Communication access for people with hearing imparied hearing*, pages 197–226, 1994.

[SW05] Cees GM Snoek and Marcel Worring. "Multimodal Video Indexing: a Review of the State-of-the-art". *Multimedia Tools and Applications*, 25:5–35, 2005.

[SWS+00] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. "Content-based Image Retrieval at the end of the Early Years". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1349–1380, December 2000.

[SWS05] Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. "Early versus Late Fusion in Semantic Video Analysis". In *Proceedings of the 13th annual ACM international conference on Multimedia (MULTIMEDIA '05)*, pages 399–402, New York, USA, 2005.

[SWY75] G. Salton, A. Wong, and C. S. Yang. "A Vector Space Model for Automatic Indexing". *Communications of the ACM*, 18:613–620, November 1975.

[SZ03] Josef Sivic and Andrew Zisserman. "Video Google: a Text Retrieval Approach to Object Matching in Videos". In *Proceedings of 9th IEEE International Conference on Computer Vision*, volume 2, pages 1470–1477, Nice, France, October 2003.

[SZ06] Josef Sivic and Andrew Zisserman. "Video Google: Efficient Visual Search of Videos". In *Toward Category-Level Object Recognition*, volume 4170, pages 127–144. 2006.

[Tam08] Sharvari Tamane. "Content based Image Retrieval using High Level Semantic Feature". In *Proceedings of the 2nd National Conference: INDIACom '08)*, February 2008.

[TIG+11] Ling-Xiang Tang, Kelly Y Itakura, Shlomo Geva, Andrew Trotman, and Yue Xu. "The Effectiveness of Cross-lingual Link Discovery". In *Proceedings of The Fourth International Workshop on Evaluating Information Access (EVIA)*, pages 1–8, 2011.

[TK09] Theodora Tsikrika and Jana Kludas. "Overview of the WikipediaMM Task at ImageCLEF 2008". In *Evaluating Systems for Multilingual and Multimodal Information Access*, pages 539–550. 2009.

[TNW08] Hung-Khoon Tan, Chong-Wah Ngo, and Xiao Wu. "Modeling video hyperlinks with hypergraph for web video reranking". In *Proceedings of the 16th ACM international conference on Multimedia*, pages 659–662, 2008.

[TTR12] Xinmei Tian, Dacheng Tao, and Yong Rui. "Sparse Transfer Learning for Interactive Video Search Reranking". *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 8:26:1–26:19, August 2012.

[TV07] Ba Tu Truong and Svetha Venkatesh. "Video Abstraction: a Systematic Review and Classification". *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 3:3, 2007.

[TYT+92] Masayuki Tani, Kimiya Yamaashi, Koichiro Tanikoshi, Masayasu Futakawa, and Shinya Tanifuji. "Object-oriented video: interaction with real-world objects through live video". In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 593–598, 1992.

[VC99] Christopher C. Vogt and Garrison W. Cottrell. "Fusion via a Linear Combination of Scores". *Information Retrieval*, 1:151–173, 1999.

[VR04] Irena Valova and Boris Rachev. "Retrieval by Color Features in Image Databases". In *Proceedings of Advances in Database and Information Systems*, pages 22–25, Budapest, Hungary, September 2004.

[VTAN13] Carles Ventura, Marcel Tella-Amo, and Xavier Giró Nieto. "UPC at Mediaeval 2013 Hyperlinking Task". In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain, October 2013.

[Wal80] Bender Walter. "animation via video disk". In *MS Thesis*. Massachusetts Institute of Technology, 1980.

[Wat89] John Watlington. "Video Finger". In *MS Thesis*. Massachusetts Institute of Technology, 1989.

[WCCS04] Yi Wu, Edward Y. Chang, Kevin Chen-Chuan Chang, and John R. Smith. "Optimal Multimodal Fusion for Multimedia Data Analysis". In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, pages 572–579, New York, USA, October 2004.

[WCL07] Che-Yen Wen, Liang-Fan Chang, and Hung-Hsin Li. "Content-based Video Retrieval with Motion Vectors and the RGB Color Model". *Forensic Science Journal*, 6:1–36, 2007.

[wik09] "An Introduction of Wikipedia". Wikimedia Foundation Inc., 2009. `http://en.wikipedia.org/wiki/Wikipedia/`.

[Wil09] Peter Wilkins. *"An Investigation into Weighted Data Fusion for Content-based Multimedia Information Retrieval"*. PhD thesis, Dublin City University, 2009.

[Wu12] Shengli Wu. "Linear Combination of Component Results in Information Retrieval". *Data Knowledge Engineering*, 71:114–126, January 2012.

[WZZL10] Shikui Wei, Yao Zhao, Zhenfeng Zhu, and Nan Liu. "Multimodal Fusion for Video Search Reranking". *IEEE Transactions on Knowledge and Data Engineering*, pages 1191–1199, August 2010.

[XZT+06] Ziyou Xiong, Xiang Sean Zhou, Qi Tian, Yong Rui, and Thomas S Huang. "Semantic Retrieval of Video". *IEEE Signal Processing Magazine*, 23:18, 2006.

[Yan06] Tao Yang. "Applications of Computational Verbs to Effective and Realtime Image Understanding". *International Journal of Computational Cognition*, 4:49–67, 2006.

[YH07] Rong Yan and Alexander G. Hauptmann. "A Review of Text and Image Retrieval Approaches for Broadcast News Video". *Information Retrieval*, 10:445–484, Oct 2007.

[YJHN07] Jun Yang, Yu-Gang Jiang, Alexander G. Hauptmann, and Chong-Wah Ngo. "Evaluating Bag-of-visual-words Representations in

Scene Classification". In *Proceedings of the International Workshop on Multimedia Information Retrieval*, pages 197–206, New York, USA, 2007.

[YJL04] J. Ye, R. Janardan, and Q. Li. "Two-dimensional Linear Discriminant Analysis". *Advances in Neural Information Processing Systems*, 17:1569–1576, 2004.

[ZCT+04] Hugo Zaragoza, Nick Craswell, Michael Taylor, Suchi Saria, and Stephen Robertson. "Microsoft Cambridge at TREC 13: Web and HARD Tracks". In *Proceedings of TREC 2004*, Gaithersburg, Maryland USA, November 2004.

[Zha08] ChengXiang Zhai. "Statistical Language Models for Information Retrieval". *Foundations and Trends in Information Retrieval*, 2:137–213, 2008.

[ZLS+06] Wujie Zheng, Jianmin Li, Zhangzhang Si, Fuzong Lin, and Bo Zhang. "Using High-level Semantic Features in Video Retrieval". In *Proceedings of Image and Video Retrieval: 5th Internatinoal Conference*, pages 370–379, Tempe, AZ, USA, July 2006.

[ZVC89] Y.T. Zhou, V. Venkateswar, and R. Chellappa. "Edge Detection and Linear Feature Extraction using a 2-D Random Field Model". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:84–95, January 1989.

List of Figures

List of Tables