Multimedia Information Seeking through Search and Hyperlinking *

Maria Eskevich
Centre for Next Generation Localisation
Dublin City University
Dublin 9, Ireland
meskevich@computing.dcu.ie

Gareth J. F. Jones
Centre for Next Generation Localisation
Dublin City University
Dublin 9, Ireland
gjones@computing.dcu.ie

Robin Aly
University of Twente
P.O. Box 217
7500AE Enschede
The Netherlands
r.aly@ewi.utwente.nl

ABSTRACT

Searching for relevant webpages and following hyperlinks to related content is a widely accepted and effective approach to information seeking on the textual web. Existing work on multimedia information retrieval has focused on search for individual relevant items or on content linking without specific attention to search results. We describe our research exploring integrated multimodal search and hyperlinking for multimedia data. Our investigation is based on the MediaEval 2012 Search and Hyperlinking task. This includes a known-item search task using the Blip10000 internet video collection, where automatically created hyperlinks link each relevant item to related items within the collection. The search test queries and link assessment for this task was generated using the Amazon Mechanical Turk crowdsourcing platform. Our investigation examines a range of alternative methods which seek to address the challenges of search and hyperlinking using multimodal approaches. The results of our experiments are used to propose a research agenda for developing effective techniques for search and hyperlinking of multimedia content.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*

General Terms

Measurement, Performance

1. INTRODUCTION

From a digital library perspective, providing users with engaging ways to interact with audiovisual content, helping them to discover, browse, navigate and search archives, is a prerequisite for opening up libraries in a way that increases their economic and/or cultural value. Users from a variety of backgrounds, such as professionals from the creative industry, journalists, students, researchers, and home users, can benefit from effective search and the interlinking of content.

Within audiovisual archives, the concept of hyper-video allows users to navigate between multimedia elements in a source content and related elements in the same content file or other multimedia sources. This can provide a means to explore additional (linked) information sources (detail-on-demand) while accessing content in a linear fashion (e.g., [12], [1],[23], [9]), or as an approach towards interactive non-linear access to video allowing users to generate narratives on-the-fly (e.g., [27], [28], [20]). In this paper, multimedia hyperlinking is addressed within an information seeking framework. This is enabled via an integrated approach that encompasses both the search process for relevant multimedia segments and the creation of links from these segments to other related multimedia segments in the archive.

To be able to investigate multimedia search and hyperlinking properly, it is crucial to develop test collections corpora and select metrics which evaluate information access in a meaningful way from the perspective of the user. Our investigation uses the test set developed for the multimedia Search and Hyperlinking using a task at MediaEval 2012 [4]. This approaches multimedia hyperlinking from an archival search scenario perspective in which a user searches for specific information within a collection of multimedia items in an audiovisual archive. The user seeks a single relevant known-item fragment of multimedia content where each content fragment in the result list is linked to other related multimedia fragments within the same collection. The tasks are based on the user-generated video collection Blip10000 for which search queries and relevance of automatically created links were created using the Amazon Mechanical Turk (AMT)[1] crowdsourcing platform. Unlike much existing work in audiovisual search this task is truly multimodal where relevance can be related to both visual and audio information streams. While our investigation builds on existing work in speech and video search, inter-item linking for multimedia collections is a much less well established, and our approaches to this task are thus more exploratory in nature. Our experimental search and hyperlinking methods explore use of audio and visual content, examining both search ef-

*This paper was written collaboratively by organizers and participants of the MediaEval 2012 Search and Hyperlinking task. Please refer to the final section of the paper for a list of the names of the other authors.

ICMR'13, April 16–20, 2013, Dallas, Texas, USA.

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

fectiveness and the behaviour of linking when using different media streams [19].

In the remainder of this paper, Section 2 reviews the context of our evaluation approach based on related work in the field, Section 3 outlines the MediaEval 2012 Search and Hyperlinking task and dataset, Section 4 provides details of the different search and hyperlinking generation techniques used in our study, Section 5 presents the results of our experiments, and Section 6 summarizes our findings and discusses potential future research directions.

2. RELATED WORK

Realisation of an integrated multimedia search and hyperlinking system requires combination of multimedia search and methods for automated creation of inter-item hyperlinks for multimedia content. Research into the effectiveness of potential methods for undertaking these tasks requires the provision of suitable test collections. This section reviews existing work in multimedia search and hyperlinking, and in the development of experimental test collections.

Our investigation of multimedia focuses on both spoken content retrieval and video search. A range of previous investigations have explored effectiveness of these tasks. The spoken document retrieval (SDR) task at TREC [8] required participants to find relevant audio recordings based on textual queries. However, the temporal nature of audio content means that playback of single retrieved documents can make information access very inefficient, and a more direct pointer to the relevant content could be beneficial. Furthermore, when searching in videos it is desirable that users can also make use the visual modality to formulate their query. The search tasks in the TRECVid workshop envision a scenario where the user poses a multimodal query and systems return shots [30]. Here, shots are treated as isolated documents and systems that return a shot just before a relevant segment do not earn any evaluation scores. Methods for searching on multimedia content have to decide on a query representation, a document representation, and a function that ranks documents according to the query using these representations. In SDR, query terms are often considered as independent keywords and the audio documents are represented by the output hypotheses of an automatic speech recognition (ASR) system. Because of the semantic gap [31], recent approaches to visual search try to first recognize semantic concepts and then relate these to the user's query. Search based on spoken and visual content is mainly approached using pattern recognition, and therefore discards the semantics of words [35].

A previous task which went some way to combining spoken and visual search is the Rich Speech Retrieval task at the MediaEval 2011 [16], which required participants to return jump-in points to indicate the start, with corresponding end-points to indicate the end of the relevant content in videos for speech acts that are specified queries. The MediaEval 2012 Search and Hyperlinking task focuses on general queries rather than speech acts, although these were still collected using the same crowdsourcing strategy [5].

Research on hyperlinking in the literature has approached this from two distinct angles: link generation that dynamically defines links[2], and hypermedia modeling that describes user behaviour and data structures at a coarse level. Link generation identifies anchors and links between both text and multimedia documents. Links are often created between text [22], or cross domain, e.g. between video collections and text [1]. In this paper, we generate links within a single video collection. The hypermedia modeling community focuses on developing models for hyperlinks in multimedia documents [10]. For example, whether links serve the purpose of creating sequential paths through a collection or provide details on the demand. The community excludes, however, the way links are defined. Our linking model focuses on topically linked video segments, rather than linking, for example, individual persons that appear in the video.

The notion of standardized tasks is not as strong in video linking research as in the search community. Research is therefore executed on individual data sets which limits repeatability. Our work is therefore among the first to carry out extensive analysis on a standardized hyperlinking task.

An important component in conducting research into multimedia search is the definition of a suitable evaluation framework. Most search evaluation tasks have utilized information needs developed by the task organizers or their associates, and with ground truth for search results created by professional annotators. This procedure has the potential disadvantage that it biases the types of investigated information needs to those imagined by the task organizers. Furthermore, the generation of ground truth in this way is expensive and may not reflect the relevance of the items to a more general population of users. By contrast, the MediaEval 2012 Search and Hyperlinking test collection was developed using crowdsourcing methods.

3. SEARCH AND HYPERLINKING TASK

The Search and Hyperlinking task at MediaEval 2012 [4] formed an initial experimental investigation of multimodal search and linking for video segments within a multimedia collection. The process was split into two intuitive sub-tasks focusing on search and hyperlinking activities. This allowed us to experiment with their combination in the task of first performing search and then forming inter-item links based on the search results. The overall task scenario is illustrated in Figure 1. User queries expressed in text, potentially enriched with visual information in multimodal queries, are entered into the search system to seek relevant segments. Since browsing through multimedia material is time-consuming, it is crucially important to start the playback of the video as close as possible to the beginning of the actual relevant segment, the so-called jump-in point. Further the user experience of browsing through the collection is enriched by the list of potential hyperlinks to the segment retrieved in the Search sub-task stage. In the following subsections we overview the description of the video dataset, and then continue with descriptions of the individual sub-tasks.

Blip10000 dataset The Blip10000 dataset created by the PetaMedia NoE [18] contains 14,838 Creative Commons videos from blip.tv, and corresponding user provided metadata. The data comprises a total of ca. 3,260 hours of data and is divided into development and test sets, of 5,288 and 9,550 videos respectively. Additionally, two transcripts were provided for all the videos in the collection, by LIMSI/Vocapia [15] and LIUM [26]. The full Blip10000 dataset used at MediaEval 2012 contains videos in different languages. How-

[2]Note that although the creation sometimes involves search, this is conceptually distinct from the first step in the search and hyperlink sub-tasks.

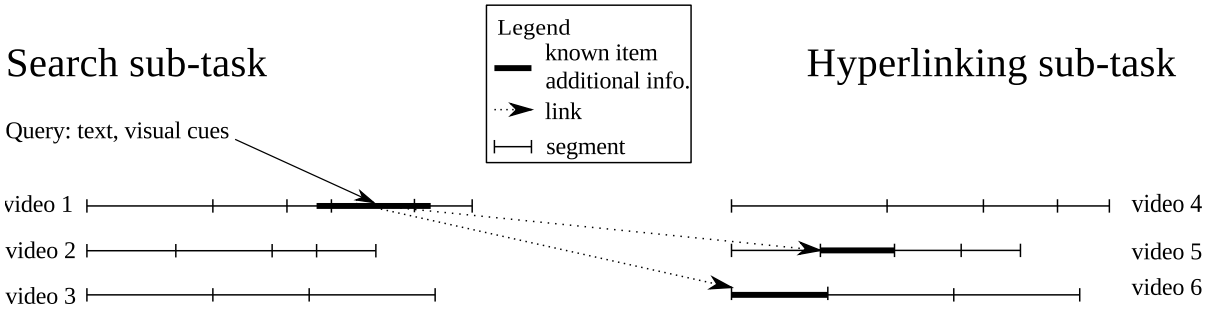

Figure 1: Overview of the search and hyperlinking task.

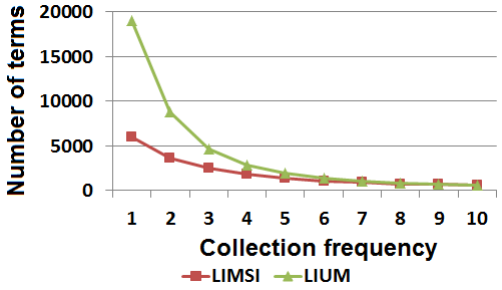

Figure 2: Number of terms with collection frequency equal to 1-10.

ever, since the main focus of our current study is to define suitable techniques and evaluation methods for a complex search and hyperlinking user experience, we wished to work only with a monolingual English language dataset. To this end, we use a subset of the test set classified as English language by LIMSI/Vocapia, and transcribed by English language versions of both ASR systems. This resulted in a corpus to 4,890 video files. The transcripts generated by the ASR systems differ in various ways, especially in the number of terms with a total collection frequency equal to 1-10, see Figure 2.

In addition to spoken transcripts, the dataset was indexed with various visual information: shot boundaries of average shot length circa 30 seconds [14] with a single visual keyframe for each shot, concept-based descriptors based on a list of 589 concepts detected using the on-the-fly video detector Visor [2], and face detection results [3]. The concept list was created by taking the tags of the Tagging task set at MediaEval 2011 and calculating a score for each tag a representing its corporeality using the algorithm described in [17]. A threshold was chosen to cut the list off at a reasonable length.

MediaEval 2012 Test Query Set The MediaEval 2012 test query set consists of 30 textual queries. These include query statements both in a natural language sentence (NLS) and in the form of a search engine request (SER) style: e.g. Query 3: textual fields: NLS: "Curtis Baylor of Allstate gives a small piece of planning advice for small business using his basic three factors.", SER: "interviews with business professionals"; multimodal features: face: "yes", colours: "dark", video content: "Chair, Man, White Shirt". The NLS query set was used for all experiments reported in this paper.

The textual information for the search sub-task was collected via crowdsourcing on the AMT platform [4], while

Table 1: Overview of the query set characteristics

query	WRR		OOV		query	WRR		OOV	
ID	limsi	lium	limsi	lium	ID	limsi	lium	limsi	lium
1	0.90	0.70	0.13	0.13	16	0.92	0.77	-	-
2	0.76	0.71	-	-	17	0.71	0.59	-	-
3	0.83	0.70	-	-	18	0.67	0.67	-	-
4	0.84	0.74	-	-	19	0.70	0.74	-	-
5	0.86	0.72	0.13	0.13	20	0.50	0.48	-	-
6	0.44	0.33	-	-	21	0.53	0.67	0.22	0.22
7	0.67	0.63	0.14	0.14	22	0.86	0.89	-	-
8	0.88	0.74	0.15	0.10	23	0.47	0.40	-	-
9	0.69	0.62	-	-	24	0.72	0.56	-	-
10	0.77	0.68	-	-	25	0.38	0.08	-	-
11	0.65	0.74	0.07	0.13	26	0.52	0.48	0.25	0.25
12	0.42	0.09	-	-	27	0.65	0.57	-	-
13	0.59	0.64	-	-	28	0.73	0.50	0.14	0.14
14	0.75	0.60	0.17	0.17	29	0.54	0.42	-	-
15	0.83	0.72	0.09	0.09	30	0.71	0.71	-	-

the multimodal features were created manually afterwards. The query set was created for a number of videos selected at random from the top 10 genre categories in the document collection. The average frequency of the query terms in the collection lexicon is relatively high (3015 and 2897 for LIMSI and LIUM respectively), although this was lower than those of the transcripts (6753 and 6342 for LIMSI and LIUM respectively). The level of out-of-vocabulary (OOV) terms is zero for 20 queries, and while for the remaining 10, it is not higher than 0.25, see Table 1. With these statistics in mind, the feature of the transcripts that has most potential to influence the retrieval performance is ASR errors. Table 1 shows the correct word recognition rate for words in relevant segments for the target segment (word recognition rate (WRR)) for each query.

Search sub-task evaluation We follow Search sub-task at MediaEval 2012, [4], in using three metrics in order to evaluate sub-task results: mean reciprocal rank (MRR), mean generalized average precision (mGAP) and mean average segment precision (MASP). Since browsing through multimedia recordings is time-consuming, we limit relevant results to a window of 60 seconds before and after the actual beginning of the relevant segment, so that retrieved segments outside of this window are considered non-relevant. Reciprocal Rank is calculated as the reciprocal value of the rank of the first correctly retrieved document [34]. mGAP [24] awards runs that not only find the relevant items earlier

in the ranked output list, but also are closer to the jump-in point of the relevant content. MASP [7] takes into account the ranking of the results and the length of both relevant and irrelevant segments that need to be listened to before reaching the relevant item.

Hyperlinking sub-task evaluation Automated inter-item multimedia hyperlinking is an open research problem for which we currently lack complete understanding of the structure and basis of the linking. Thus we investigated link creation using different approaches to seek a better understanding of the different types of links that can be formed using different linking methods and query modalities. Similar to the creation of standard ad hoc search test collections [33], a posteriori evaluation strategy was used to carry out relevance assessment of proposed links. Since we had modified the document set and carried out new runs, we were unable to use the MediaEval 2012 Search and Hyperlinking link relevance data, and thus repeated the assessment stage for our new runs. The top 10 ranked segments from the linking runs were evaluated using the AMT platform. The qrel set collected for each method ("run qrel") enables us to assess how effective it is for finding links that based the features used in this run (e.g. audio content, visual quality etc). The full set of all relevant segments gathered for all runs was formed into a single unified qrel file ("unified qrel"). Since the alternative approaches to hyperlink creation may use different segment definitions, these may result in overlap of the relevant links. In this case we combined these into a single relevant link after manual assessment. This unified list of manually assessed links was used to calculate mean average precision (MAP) across all the methods. Analyzing the results in this way enables us to examine the diversity of links proposed by each method.

As this approach to evaluation of the hyperlinking results is experimental, crowdsource workers were not given specific details (audio or visual features) to define videos relatedness. Within AMT the task and its setting is referred to as a 'Human Intelligence Task' or HIT. The following options were given in the HIT with a separate field for explanation of the choice and comments from the worker.

1. The video segments are totally unrelated to one another. The second video is something new and different from the first.
2. The video segments are related, they have the same topic or focus, though they give different information.
3. The video segments are related, but they give a different perspective or view on the same information.
4. The video segments are basically the same. If I saw one, I really wouldn't need to see the other.

If options 2, 3, or 4, or their combinations were selected, the passage being examined was considered relevant.

The questions requiring detailed answers were included to prevent low quality work being included in the relevance assessment results. Overall we approved 3165 HITs, and 1759 were rejected because the answers were not correct. With the price of the HIT equal to 0.09$, the overall cost of the evaluation was equal to 314$.

4. APPROACHES TO SEARCH AND HYPERLINKING

In this section we introduce the different approaches to addressing the sub-tasks. Exact details of each run are given in Section 5.

Table 2: Overview of Search Sub-Task Types of Runs

Segmentation Method	Use of Metadata		Retrieval Model
	–	+	
Sentence	*S_bm25	*SM_bm25	BM25
Shot	Sh_tf-idf	ShM_tf-idf	TF-IDF
Speech Segment	*SpS_bm25	*SpSM_bm25	BM25
Time + Pause	TP_bm25		BM25
	TP_lm		LM
	TP_tf-idf		TF-IDF
Time + Overlap	TOv_bm25	TOvM_bm25	BM25
	TOv_lm	TOvM_lm	LM
	TOv_tf-idf	TOvM_tf-idf	TF-IDF

4.1 Search Sub-task

The main goal of the search sub-task is to find the exact topically coherent passage in the video collection relevant to the given query. Failure to achieve this task can negative impact on automatic hyperlinking performance and the overall user experience. According to the problem setup we have two types of queries: textual and multimodal (textual information is enriched with video content information). In this section we describe our approaches with respect to the following aspects: segmentation method, use of metadata, and use of different retrieval models. The overall list of the runs and naming convention is shown in Table 2. First we put the segmentation method with the first letters of the name, 'M' distinguishes runs that use the metadata, and then the retrieval model used is named (for example, Shot segmentation using Metadata and the TF-IDF IR model is named 'ShM_tf_idf'). Runs with labels starting with an asterix (*) use both textual and multimodal queries.

Video Segmentation Videos within the collection can vary in length and number of topics covered. Therefore they need to be segmented into shorter passages which can be handled as "documents" in the traditional retrieval setup. Segmentation can be based on audio and visual features: the former include time information (runs types 'TP', 'TOv') and various sub-structures found in the ASR transcript corresponding to the video (sentences (run type 'S') and speech segments (run type 'SpS'), pauses (run type 'TP')), the latter uses shot segmentation results (runs 'Sh').

Sentences and speech segments boundaries are taken from the ASR transcript. Time stamps in the ASR transcript allow addition of boundaries at pauses in the speech (a distance between two words of ≥ 0.5 seconds is considered to be a pause), segmentation into the range of different segment lengths with window overlap of different size. Another source for potential segment boundaries based on the video channel are the shot boundaries.

Retrieval models We experimented with three different retrieval models: BM25 [25], Hiemstra's Language Model [11] with default parameters, and TF-IDF. We used the Terrier[3] IR system for runs based on time segmentation (types 'T(M)', 'TOv(M)'), and the Lucene IR system for runs based on sentences and speech segments (types 'S(M)', 'SpS(M)') In all methods standard stop words removal and Porter stemming was carried out.

Metadata exploitation Since ASR transcripts contain

[3]www.terrier.org

errors, and may not contain all relevant information for the videos, any available metadata becomes potentially useful to compensate for this information loss. When available the metadata is used, though only provided at the video level rather than for a specific video segment.

4.2 Hyperlinking Methods

Since approaches to the hyperlinking sub-task are not as well established as those for the search sub-task, multiple approaches were implemented in order to get a broader perspective on potential solutions. We used all modalities of the provided dataset: metadata, information based on the audio (transcript words, transcript sentences and segmented speech units) and video (shots boundaries, visual concepts) processing. Additionally an enrichment tool (DBpedia Spotlight [21]) was used to identify named entities (NE) in the transcript and link these to Linked Data resources. Visual features were also enriched by clustering the information from SIFT descriptors into 500 clusters corresponding to visual words, as described in [29].

All approaches use the following main processing stages:

1. Retrieval of potential links to videos or video segments, when the retrieval units might be different from the target units (the whole video instead of the shot).
2. Re-ranking of the returned potential links.
3. Segment extraction. In cases where the retrieval units do not correspond to the target units.

The difference in implementation and ordering of these stages accounts for the differences in the results of the reported runs. Table 3, gives an overview of the representation of the data and retrieval techniques used for the runs. As in case of search sub-task, determining the boundaries of the target segment of a video link is important because they are shown to the user. When naming the runs we start with the first letters of the target segment (even though at retrieval stage those segments might be part of larger units), then after an underscore we add the types of the data used and conclude with adding the letter 'M' in case of use of metadata, for example the run that target to find the segments of shot size and uses speech transcript, visual words, and metadata is called 'Sh_TVWM'. In the next sections we describe run details of retrieval models, reranking strategies and segment extraction.

4.2.1 Retrieval Methods for Hyperlinking

Different retrieval techniques were used to identify and rank potential segments for hyperlinking. Videos and queries were represented as TF-IDF vectors with alternative compositions consisting of: transcript words only ('TrT_T'); transcript words and metadata ('TrT_TM' and 'Sh_TM',). Another approach was to have several TF-IDF vectors for each representation and to combine the ranking results later in the process, this approach is labeled run 'Sh_TVWM' and combines two vectors: transcript words and metadata, and visual words. In run 'Sh_TNE(M)', the TF-IDF vector uses Named Entities (NE) instead of words: a new weight called *TF-IS* (*term-frequency-inverse-support*) is used which depends on the number of URLs linking to this NE in an RDF graph of DBpedia Spotlight; and the metadata tags are ranked using the Jaccard similarity measure. The 'SpS_VCM' run uses speech segment boundaries (but not the transcript itself), and represents the videos by vectors of 508 visual concepts associated with their confidence scores provided with the collection, with a utilizes Euclidean distance to rank the resulting hyperlinks.

4.2.2 Ranking Strategies

For each of the runs, except runs 'SpS_VCM' and 'Sh_TM', a post-processing step was performed to re-rank the result list obtained using the retrieval methods described in the previous section. Runs 'TrT_T(M)' used a re-ranking strategy in order to favour videos that were part of the same series of videos in the collection as that of the query. To do so, the scores of each such video were artificially increased to make them appear at the top of the list. The 10 videos associated with the highest scores were then extracted from this modified list at potential links. Run'Sh_TVWM' had two variations of re-ranking scheme: a score-based and a rank-based fusion that combined the similarity computations based on textual and visual information. In run 'Sh_TNE(M)', a short list of results was extracted for each comparison scheme (using words, NEs or tags) using a different threshold for each one. Scores of videos that appear in several of these short lists were added together and the final list of results obtained by selecting the 40 videos associated with the highest scores.

4.2.3 Segment Extraction Strategies

Our aim is to link the videos on the level of extracted segments, so users can be directed closer to the part of the video they are interested in. We experimented with two ways of retrieving segments: perform retrieval and re-rank directly on a set of segments, or run retrieval and re-rank on the entire video, and extract segments from the top results afterwards.

Segmentation relies on either audio or visual information. Runs 'TrT_T(M)' and 'SpS_VCM' use the transcript information. Run 'SpS_VCM' uses speech segment boundaries already available at the indexing stage. Run 'TrT_T(M)' extracts segments based on the topical relevance of the content of the complete video (if the video contained several topics[4]) subsequent to retrieval only for videos ranked in top positions. If segment extraction is not applied, a sliding window of 40 words is used to identify the segment in the transcript containing the largest number of content words in the query. Runs 'Sh_TM' and 'Sh_VWM' used visual segments extracted using visual analysis prior to retrieval and combination ranking using multimodal fusion, whereas run 'Sh_TNE(M)' segmented the retrieved video into shot segments only at the final stage.

5. EXPERIMENTS

In this section we give details of the exact runs and results of our experiments.

5.1 Search Sub-Task

Table 2 introduced in Section 4.1 shows the general types of the runs. Since the dataset provides two types of transcripts, some of the runs were carried out for both transcript types.

Sentence and speaker level segmentations output by the ASR system and visual shot boundaries were provided with

[4]To decide if a transcript contained several topics, a topic segmentation algorithm [32] parameterized to over-segment documents was applied. If the number of topic segments returned is small (less than 10), it was assumed that the transcript contained only one topic.

Table 3: Overview of Hyperlinking Types of Runs (BoW: bag of words, NE: Named entities, M - use of Metadata)

Run Type	Anchor Representation	Target Segment	Representation				Ranking Function
			Metadata (BoW)		Speech	Visual	
TrT_T(M)	QueryBoW	Transcript Topics (TrT)	-	+	BoW		BM25
Sh_TM	QueryBoW	Shots (Sh)		+	BoW		TF-IDF
Sh_TVWM	QueryBoW Keyframe (Sh)	Shots		+	BoW	Visual Words (VW)	TF-IDF
Sh_TNE(M)	QueryBoW+NE	Shots (Sh)	-	+	BoW+NE		TF-IDF
SpS_VCM	Keyframes concepts cores	Speech segment (SpS)				Visual Concepts (VC)	Euclid. Dist.

Table 4: MRR for segmentation using varying segment (60-180) and overlap (10-180) lengths

Segment Size (sec)	Overlap Size (sec)						
	10	30	60	90	120	150	180
LIMSI							
60	0.393	0.392	**0.408**	—	—	—	—
90	0.340	**0.429**	0.316	0.344	—	—	—
120	**0.337**	0.323	0.312	0.337	0.292	—	—
150	0.282	**0.305**	0.286	0.280	0.224	0.260	—
180	0.215	0.251	0.264	**0.269**	0.209	0.245	0.260
LIUM							
60	**0.358**	0.348	0.327	—	—	—	—
90	0.326	**0.349**	0.235	0.327	—	—	—
120	**0.322**	0.294	0.306	0.288	0.270	—	—
150	0.227	**0.281**	0.267	0.269	0.202	0.243	—
180	0.161	0.215	0.249	**0.276**	0.184	0.205	0.252

the dataset. Runs using fixed length time segmentation enabled us to vary the duration of segments in a controlled manner. Previous work on SDR has shown good performance for systems that retrieve segments with boundaries close to the actual relevant data. This can be assisted for fixed length segments by the incorporation of audio pauses ($\geq$ 0.5 secs) as additional boundaries to divide speech segments, and a sliding window strategy with further filtering [6]. As a simple filtering technique to keep only the first segment after removing all overlapping segments further down the result list proved to be helpful to improve search retrieval results in [13]. We apply this filtering method to our time based segmentation runs. In order to choose the optimal values for time segmentation we carried out preliminary runs using different segment and segment overlap values. In these runs we use the Hiemstra language model (LM) with $\lambda = 0.15$, see Table 4 for MRR results. Using segment length equal to 90 seconds with an overlap of 30 seconds received the highest score amongst all runs based on LIMSI and LIUM transcripts, therefore we used these values for all time-based runs.

Table 5 shows results for all the runs. In general time based segmentation with overlapping windows shows the highest MRR values, meaning that these methods get the relevant content at the higher ranks. Amongst all runs 'TOvM_lm' achieves the highest score for both type of transcripts. mGAP values incorporating the distance to the beginning of the retrieved segment from the actual relevant jump-in point show less difference between the methods, meaning that some of the segments retrieved by time-based segmentation at higher ranks start further from the jump-in point, than the shorter segments of the other runs being retrieved further down the list. The influence of segment length also affects MASP performance, where shorter segments ('S(M)_bm25') perform better than all other segmentation methods for both types of transcripts.

When we compare the results for the same methods using different types of transcripts, the segmentation methods that use shorter segments have better scores when LIUM transcript is used, whereas time-based segmentation performs better on the LIMSI transcript. Amongst methods that produce short segments ('S(M)', 'Sh(M)','SpS') and use LIMSI transcripts, the addition of metadata increases results in all cases expect for shot based segmentation. Addition of metadata increases the results for sentence and speech segment based runs for LIMSI transcripts (S_bm25 vs SM_bm25, SpS_bm25 vs SpSM_bm25). However for time based segmentation, it decreases results for runs using the TF-IDF retrieval model (10a vs 13a). TF-IDF and BM-25 perform better than LM when metadata is used, however when time segmentation is combined with pause information, LM outperforms BM25 and TF-IDF.

When we analyze run performance depending on query type, the difference between segmentation methods becomes more obvious. For queries with WRR equal or less than 0.55 (queries 6, 12, 23, 25, 26, 29) runs based on short segmentation units (sentences, shots, speech segments) get higher MRR values than time-based segmentation with overlapping windows (e.g., in the extreme cases it is equal to 1.0 vs 0.0 for queries 6, 25 for time segmentation and short units segmentation respectively). This is due to the fact that in longer segments the errors in ASR recognition cause this type of behaviour.

Queries containing terms that are not present in the collection lexicon (OOV terms) get better results for longer segments, and especially for runs that use metadata (e.g. only TOvM runs retrieve the relevant segment for queries 1, 8, 11; metadata increased the performance for all types of runs for queries 5, 7, 14, 15, 26, 28). This is due to the fact that metadata can contain the missing OOV words, and the longer segments contain more context that relates the segment to the terms appearing in the query that are present in the collection lexicon.

Addition of visual information in the multimodal queries decreases performance for both types of transcript, although more significantly for the runs based on LIUM transcripts. This may be caused by the fact that the queries are primarily expressed in terms of audio content, thus the video stream is less relevant for search.

Table 5: Evaluation metrics for Search sub-task.

Segmentation Type	LIMSI			LIUM		
	MRR	mGAP	MASP	MRR	mGAP	MASP
S_bm25	0.127	0.097	0.167	0.349	0.258	**0.213**
SM_bm25	0.277	0.206	**0.240**	0.349	0.258	**0.213**
Sh_tf-idf	0.187	0.120	0.066	0.275	0.158	0.032
ShM_tf-idf	0.158	0.096	0.055	0.269	0.154	0.029
SpS_bm25	0.235	0.188	0.140			
SpSM_bm25	0.363	0.264	0.220			
TP_bm25	0.212	0.150	0.082	0.164	0.128	0.128
TP_lm	0.336	0.236	0.097	0.318	0.237	0.181
TP_tf-idf	0.212	0.150	0.082	0.162	0.126	0.127
TOv_bm25	0.436	0.284	0.099	0.390	0.248	0.076
TOv_lm	0.364	0.241	0.078	0.355	0.232	0.066
TOv_tf-idf	0.414	0.276	0.085	0.405	0.260	0.078
TOvM_bm25	0.423	0.251	0.102	0.429	0.238	0.091
TOvM_lm	**0.470**	**0.290**	0.123	**0.449**	**0.250**	0.102
TOvM_tf-idf	0.428	0.256	0.103	0.418	0.239	0.087
*S_bm25	0.126	0.096	0.144	0.080	0.044	0.046
*SM_bm25	0.104	0.058	0.071			
*SpS_bm25	0.196	0.168	0.102			
*SpSM_bm25	0.121	0.072	0.070			

Table 6: MAP results for Hyperlinking sub-task.

Run Type	MAP using			
	run qrel		unified qrel	
	LIMSI	LIUM	LIMSI	LIUM
TrT_T	0.251	0.222	0.156	0.137
TrT_TM	0.334		**0.208**	
TrT_T_reranking1	0.346		0.206	
TrT_T_reranking2	0.315		0.192	
Sh_TM	0.254	0.312	0.099	0.134
Sh_TVWM_VisualReranking	0.194	0.245	0.076	0.105
Sh_TVWM_RankReranking	0.228	0.258	0.091	0.111
Sh_TNE	0.088	0.083	0.016	0.020
Sh_TNEM	0.088	0.083	0.016	0.020
SpS_VS	**0.404**		0.055	
*Sh_TNE	0.071	0.046	0.018	0.012
*Sh_TNEM	0.071	0.046	0.018	0.012
TOv_lm+Sh_TM			0.004	0.004
TOv_lm+ Sh_TVWM_RankReranking			0.003	0.002
TOvM_lm_+Sh_TM			0.015	0.019
TOvM_lm+ Sh_TVWM_RankReranking			0.014	0.018

5.2 Hyperlinking Sub-Task

As described in Section 3 the relevance assessment for the hypothesized hyperlinks was conducted after completing the runs. Table 6 shows the results of the runs using different transcript types, with MAP values for two types of qrels. The run using only visual features for retrieval (SpS_VS) has the highest MAP on its "run" qrel and the lowest on the "unified" qrel, this means that it is oriented to retrieve links based on video stream features, but cannot retrieve the other types of links. Use of visual words in combination with the transcript and metadata (runs 'Sh_TVWM_VisualReranking' and 'Sh_TVWM_RankReranking') decreases the result in comparison with the simple run based only on transcript and metadata ('Sh_TM'). We assume that visual features used are too low level to improve the results.

The results here cannot be considered definitive, further development of the evaluation set may change the results. For example it is notable that runs have very diverse qrel sets, so further runs may significantly change the unified qrels. Also we did not explicitly ask the assessors to pay attention to audio or visual features for each case, therefore it is hard to determine cases where visual or audio features have more importance.

5.3 Bridge between Search and Hyperlinking

Our target is to create links for the search output results, therefore in a real-life scenario we would have to use the segments extracted automatically as the link sources, and not the ones with the perfect boundaries. Therefore we created two additional types of runs: one using the same units in both search and hyperlinking sub-tasks (*Sh_TNE and Sh_TNEM), and the other using the two best results from the Search sub-task ('TOv' and 'TOvM'), and carry out hyperlink creation using the ASR transcripts in the source segment, rather than the manual ones used in the Hyperlinking task (runs TrT(M))). The second set of results clearly demonstrates that even a slight difference in Search sub-task output effects hyperlink creation. This motivates further work on refinement of the output of the Search sub-task.

6. CONCLUSIONS AND FUTURE WORK

This paper has described our investigation into a scenario modeling multimodal search and automated hyperlinking of multimedia content, for a situation where a user wishes to search for a remembered specific item in a collection, but does not remember where it is located, and subsequently wants to follow hyperlinks to related videos in the collection that would enrich their browsing experience. We are among the first to work on both sub-tasks of multimedia search and hyperlinking creation using multimodal aspects of the collection with one scenario in mind, using a standard benchmark collection as the dataset.

We explored the space of possible approaches to the proposed search sub-task by varying segment length, use of metadata, and different retrieval models, in order to better understand and address the search sub-task for textual and multimodal queries. We investigated various use of audio and video features for the hyperlinking sub-task, and carried out several combination runs using search sub-task output and hyperlinking methods.

Use of visual features for the search sub-task impacted negatively on the results: multimodal queries received lower scores for the same methods, and shot segmentation did not outperform time or transcript based results.

Our proposed pooled strategy for evaluation of the hyperlinking sub-task did not include the distinction at the stage of relevance assessment whether the audio or the visual features were the basis for the decision. However, it did allow us to assess separately the performance of each individual method of hyperlink creation and compare this with a unified list of relevant data gathered from all runs. This showed that runs using visual features fail to hypothesize hyperlinks that are relevant because of the audio content. This other type of links can be found using methods based based on

transcript and metadata processing. We tried to combine this method with low-level visual features. However this again resulted in a decrease in results. We can thus conclude that exploitation of visual features is a challenging issue for our future work.

The results of combination search and hyperlinking runs that reflect the full real-life scenario of search and link browsing showed the importance of appropriate search sub-task output.

7. ACKNOWLEDGMENTS

This work was supported by Science Foundation Ireland (Grant 08/RFP/CMS1677) Research Frontiers Programme 2008 and (Grant 07/CE/I1142) as part of the Centre for NExt Generation Localisation (CNGL) project at DCU; co-funded by European Commission's Seventh Framework Programme (FP7) as part of the AXES project (ICT-269980) and (FP7-269980); by the Dutch COMMIT program; by Ghent University, iMinds, the IWT Flanders, the FWO-Flanders, and the European Union, in the context of the iMinds project SMIF; by the Czech Science Foundation (grant no. P103/12/G084).

8. ADDITIONAL AUTHORS

Roeland J.F. Ordelman (University of Twente, The Netherlands, email: ordelman@ewi.utwente.nl), Shu Chen (Dublin City University, Ireland, email: shu.chen4@mail.dcu.ie), Danish Nadeem (University of Twente, The Netherlands, email: d.nadeem@utwente.nl), Camille Guinaudeau, Guillaume Gravier, Pascale Sébillot (IRISA/University of Rennes 1, IRISA/CNRS, IRISA/INSA, France, email: {cguinaud, ggravier, psebillo}@irisa.fr), Tom de Nies, Pedro Debevere, Rik Van de Walle (Ghent University, iMinds, MMLab, Belgium, email: {tom.denies, pedro.debevere, rik.vandewalle}@ugent.be), Petra Galuščáková, Pavel Pecina (Charles University in Prague, Czech Republic, email: {galuscakova, pecina}@ufal.mff.cuni.cz), Martha Larson (Delft University of Technology, Delft, The Netherlands, email: m.a.larson@tudelft.nl)

9. REFERENCES

[1] M Bron, B Huurnink, and M de Rijke. Linking archives using document enrichment and term selection. In *Proceedings of TPDL 2011*, pages 2357–2360, 2011.

[2] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding method. In *Proceedings of BMVC 2011*, 2011.

[3] R.G. Cinbis, Jakob Verbeek, and Cordelia Schmid. Unsupervised Metric Learning for Face Identification in TV Video. In *Proceedings of ICCV 2011*, Barcelona, Spain, 2011.

[4] M. Eskevich, G.J. F. Jones, S. Chen, R. Aly, R.J.F. Ordelman, and M. Larson. Search and Hyperlinking Task at Mediaeval 2012. In *MediaEval*, volume 927 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2012.

[5] M. Eskevich, G.J.F. Jones, M. Larson, and R.J.F. Ordelman. Creating a data collection for evaluating rich speech retrieval. In *Proceedings of LREC 2012*, Istanbul, Turkey, 2012.

[6] M. Eskevich, G.J.F. Jones, M. Larson, C. Wartena, R. Aly, T. Verschoor, and R.J.F. Ordelman. Comparing retrieval effectiveness of alternative content segmentation methods for internet video search. In *Proeedings of CBMI 2012*, 2012.

[7] M. Eskevich, W. Magdy, and G.J.F. Jones. New metrics for meaningful evaluation of informally structured speech retrieval. In *Proceedings of ECIR 2012*, pages 170–181, 2012.

[8] J.S. Garofolo, C.G.P. Auzanne, and E.M. Voorhees. The TREC spoken document retrieval track: A success story. In *Proceedings of RIAO 2000*, pages 1–8, 2000.

[9] A. Girgensohn, L. Wilcox, F. Shipman, and S. Bly. Designing affordances for the navigation of detail-on-demand hypervideo. In *Proceedings of AVI 2004*, pages 290–297. ACM, 2004.

[10] L. Hardman. *Modelling and authoring hypermedia documents.* PhD thesis, Universiteit Amsterdam, 1998.

[11] D. Hiemstra. *Using language models for information retrieval.* PhD thesis, University of Twente, 2001.

[12] P. Hoffmann, T. Kochems, and M. Herczeg. HyLive: Hypervideo-Authoring for Live Television. In *Changing Television Environments*, pages 51–60. Springer, 2008.

[13] T. Kaneko, T. Takigami, and T. Akiba. STD based on hough transform and SDR using STD results: Experiments at NTCIR-9 SpokenDoc. In *Proceedings of Ninth NTCIR Workshop Meeting*, 2011.

[14] P. Kelm, S. Schmiedeke, and T. Sikora. Feature-based Video Key Frame Extraction for low Quality Video Sequences. In *Proceedings of WIAMIS 2009*.

[15] Lori Lamel and Jean-Luc Gauvain. Speech processing for audio indexing. In *Advances in Natural Language Processing*, volume 5221 of *LNCS*, pages 4–15. 2008.

[16] M. Larson, M. Eskevich, R. Ordelman, C. Kofler, S. Schmiedeke, and G. J. F. Jones. Overview of MediaEval 2011 Rich Speech Retrieval Task and Genre Tagging Task. In *MediaEval 2011 Workshop*, Pisa, Italy, 2011.

[17] M. Larson, C. Kofler, and A. Hanjalic. Reading between the tags to predict real-world size-class for visually depicted objects in images. In *Proceedings of ACM MM*, 2011.

[18] M. Larson, M. Soleymani, M. Eskevich, P. Serdyukov, R.J.F. Ordelman, and G. J. F. Jones. The community and the crowd: Multimedia benchmark dataset development. *IEEE MultiMedia*, 19(3):15, 2012.

[19] M.A. Larson, S. Schmiedeke, P. Kelm, A. Rae, V. Mezaris, T. Piatrik, M. Soleymani, F. Metze, and G.J.F. Jones, editors. *Working Notes Proceedings of the MediaEval 2012 Workshop*, volume 927 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2012.

[20] B. Meixner, K. Matusik, C. Grill, and H. Kosch. Towards an easy to use authoring tool for interactive non-linear video. *Multimedia Tools and Applications*, pages 1–26, 2012.

[21] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*, 2011.

[22] D. Milne and I.H. Witten. Learning to link with wikipedia. In *Proceeding of CIKM 2008*, pages 509–518. ACM, 2008.

[23] J. Morang, R.J.F. Ordelman, F.M.G. de Jong, and A.J. van Hessen. InfoLink: analysis of Dutch broadcast news and cross-media browsing. In *Proceedings of ICME 2005*, Los Alamitos, 2005.

[24] P. Pecina, P. Hoffmannova, G. J. F. Jones, Y. Zhang, and D. W. Oard. Overview of the CLEF 2007 cross-language speech retrieval track. In *Proceedings of CLEF 2007*, pages 674–686, 2007.

[25] S. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. In *Proceedings of ACM CIKM 2004*, 2004.

[26] A. Rousseau, F. Bougares, P. Deléglise, H. Schwenk, and Y. Estèv. Lium's systems for the iwslt 2011 speech translation tasks. In *Proceedings of IWSLT 2011*, 2011.

[27] I. Sawhney, N. and Balcom, D. and Smith. Authoring and navigating video in space and time. *MultiMedia, IEEE*, 4(4):30–39, 1997.

[28] F. Shipman, A. Girgensohn, and L. Wilcox. Authoring, viewing, and generating hypervideo: An overview of Hyper-Hitchcock. *ACM Trans. Multimedia Comput. Commun. Appl.*, (2):15:1—-15:19, 2008.

[29] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *Proceedings of ICCV 2003*, pages 1470 –1477 vol.2, 2003.

[30] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *Proceedings of MIR 2006*, Santa Barbara, California, USA, 2006.

[31] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000.

[32] M. Utiyama and H. Isahara. A statistical model for domain-independent text segmentation. In *Proceedings of ACL 2001*.

[33] E. Voorhees, D.K. Harman, National Institute of Standards, and Technology (US). *TREC: Experiment and evaluation in information retrieval*. MIT press USA, 2005.

[34] E.M. Voorhees. The TREC-8 Question Answering Track Report. In *Proceedings of TREC-8*, pages 77–82, 1999.

[35] R. Yan. *Probabilistic Models for Combining Diverse Knowledge Sources in Multimedia Retrieval*. PhD thesis, Carnegie Mellon University, 2006.