An Investigation into Feature Effectiveness for Multimedia Hyperlinking

Shu Chen[1,2], Maria Eskevich[2], Gareth J. F. Jones[2], and Noel E. O'Connor[1]

[1] INSIGHT Centre for Data Analytics
Dublin City University, Dublin 9, Dublin, Ireland
shu.chen4@mail.dcu.ie, noel.oconnor@dcu.ie,
[2] CNGL Centre for Global Intelligent Content, School of Computing
Dublin City University, Dublin 9, Dublin, Ireland
{meskevich, gjones}@computing.dcu.ie ,

Abstract. The increasing amount of archival multimedia content available online is creating increasing opportunities for users who are interested in exploratory search behaviour such as browsing. The user experience with online collections could therefore be improved by enabling navigation and recommendation within multimedia archives, which can be supported by allowing a user to follow a set of hyperlinks created within or across documents. The main goal of this study is to compare the performance of different multimedia features for automatic hyperlink generation. In our work we construct multimedia hyperlinks by indexing and searching textual and visual features extracted from the blip.tv dataset. A user-driven evaluation strategy is then proposed by applying the Amazon Mechanical Turk (AMT) crowdsourcing platform, since we believe that AMT workers represent a good example of "real world" users. We conclude that textual features exhibit better performance than visual features for multimedia hyperlink construction. In general, a combination of ASR transcripts and metadata provides the best results.

Keywords: Multimedia, Hyperlinking, Crowdsourcing, Information Retrieval

1 Introduction

Fully realizing the value of the increasing amount of multimedia archival content available online requires users to engage in exploratory search behaviour to find materials which may be of interest to them. Users are increasingly not as interested in simply re-finding information contained in known-items as in the past – they wish to explore unfamiliar archives of multimedia content. This user activity can be supported by providing a set of hyperlinks within or across documents within an archive or archives. Hyperlinks should be constructed based on the semantic information described by text or visual contents of the archive. A rich and semantically meaningful set of hyperlinks can potentially improve the user experience by enabling navigation and recommendation.

Since the requirement for hyperlinks arises from the needs and interests of users, it follows that an investigation of hyperlink generation in multimedia data collections should be user-driven. Workers engaged by crowdsourcing platforms represent a good example of real potential users of multimedia browsing applications because they fit the profile of experienced Internet users, and they are able to perform relevance assessment [1]. Thus, investigation into multimedia search and hyperlinking can be based on available research video collections, whilst workers from a crowdsourcing platform can play the role of the users that help us to define which multimedia features can contribute to effective hyperlink construction.

The main goal of this paper is to compare the performance of different multimedia features for automatic hyperlink generation. State-of-the-art multimedia retrieval techniques are used to create hyperlinks within the video collection automatically. These techniques determine the relatedness between source video segments, termed *anchors* and target video segments. Workers from the crowdsourcing platform act as real-time users of a multimedia retrieval system. They are asked to watch a query video segment (anchor) and a potentially related video segment extracted by our automatic hyperlink generation process, and provide feedback on whether those segments are indeed related.

The paper is structured as follows: Section 2 overviews related work on multimedia hyperlinking and crowdsourcing techniques. Section 3 presents the design of our hyperlinking strategy, including data description and hyperlinking algorithm description. Section 4 provides experimental results and the details of user feedback. Section 5 concludes the paper and comments on our further work plans.

2 Related Work

There are a number of examples of the utilisation of links to automatically augment textual information for research or commercial purposes. Examples of this approach include the Smart-Tag service developed by Microsoft which aims to construct links between web pages or Google AutoLink which links street addresses or ISBNs to related internet resources. However, early linking systems caused numerous controversies, since many people expressed concerns that hyperlinks were being "surreptitiously" modified for commercial purposes [16]. Hyperlinking research has gradually become oriented towards non-profit data collections, such as Wikipedia. In [15] the authors presented a link creation system "Wikify!" based on Wikpedia resources. This system combined automatic document keyword extraction and word sense disambiguation to provide a rich text annotation service. The authors in [16] presented an alternative strategy using machine learning to identify significant terms within unstructured documents and enrich them with links to the appropriate Wikipedia articles. The principle of relatedness was used to exclude the situation where links were determined by a rare sense of a word, according to the incoming and outgoing links to the current Wikipedia document. In [2], the authors presented work on link-

ing multimedia resources for unskilled users, defined as exhibiting exploratory behaviour in [3]. Hyperlinking research has also appeared in the area of digital libraries focussing on news, multimedia and cultural heritage archives. The linking task was redefined 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 [2].

The VideoCLEF 2009 tasks included a multimedia hyperlinking task which required participants to find related resources across languages This was based on linking videos to material on the same subject in a different language [12]. The MediaEval 2012 benchmark campaign introduced the Search and Hyperlinking task as a Brave New Task. The idea of the task was to connect two activities in one framework, a video segment search task was combined with a separate subtask which used relevant segments as anchors from which links to other video segments should be formed within the Hyperlinking sub-task [7]. The similarity between query and target anchors was determined by participants using either of both of textual information from metadata or spoken transcripts, and visual content within shot segments [6].

Evaluation of hyperlinking systems can be carried out either based on groundtruth data collections or based on human evaluation of results. The cross-lingual hyperlinking task ay NTCIR-10 in 2012 provided two evaluation instances – automatic evaluation against queries created from the Wikipedia groundtruth and manual assessment of results [20]. In our opinion, the complexity of video content means that the evaluation of multimedia hyperlinking is best served by manual evaluation based on human judgements. Crowdsourcing is a method of having people do things that we might otherwise consider assigning to a computing device to calculate automatically [9]. As such, it offers scalable pools of workers available on-demand to offer a flexible means of gathering human judgements as needed to evaluate hyperlink construction.

3 Experimental Design

This section describes the data used for our evaluation of multimedia hyperlinking, and the strategy and features used to form these links in this study.

3.1 Data Description

The dataset used for the experiment consists of semi-professional videos uploaded to the Internet video sharing platform Blip.tv[3]. These videos are gathered into the blip10000 collection [18]. Following the setup of the Search and Hyperlinking task at MediaEval 2012, for our hyperlinking experiments, we make use of the test set in the collection that contains 9,550 videos and has a runtime of 2,125 hours [7]. The dataset comprises metadata that was manually assigned to each video by the user who uploaded it. The shot boundary of each episode was automatically created by TU Berlin [10]. The number of shot segments is 42,000 with

[3] http://www.blip.tv/

an average duration of 30 seconds. Each shot segment has an associated keyframe extracted from the middle of the shot. To analyze spoken information, two automatic speech recognition (ASR) transcripts are provided by LIMSI/Vocapia Research[4] and LIUM Research team[5]. Spoken transcripts from LIMSI/Vocapia were created by first using a language identification detector (LID) and then running an appropriate ASRS system [11]. The LIUM system is based on the CMU Sphinx project [17]. In our investigation we use the 1-best ASR transcription hypotheses only.

We define a hyperlink as a constructed link between two video segments within the collection, one a query anchor, the other a target segment. Each anchor or segment contains the start and end time within the video, and corresponding audio and visual channels. A query anchor simulates a user's request while browsing using a hyperlinking system. All 30 query anchors used in our hyperlinking system were taken from the test set of the MediaEval 2012[6] Search and Hyperlinking task. Each query contains a corresponding filename and a duration to describe the video segment boundary of the current query. Each query is associated with a piece of text description extracted from the corresponding LIMSI or LIUM transcripts. All spoken words within the video segment boundaries are included. To represent the visual content of query anchors, a keyframe located at the middle of an anchor shot is extracted by using *ffmpeg*[7].

A target segment is a section of video within the collection which we assume to be of interest to users, and that would enrich their browsing experience. In our hyperlinking system, a target segment is based on automatically detected video shots. Since the shots vary in length, we define the length of a target segment to be between 90 and 120 seconds, based on previous crowdsourcing experience in MediaEval 2012 [5]. Thus, any shot shorter than 90 seconds is expanded by combining it with nearby shots, while any shot longer than 120 seconds is cut into a segment of 120 seconds from its start point. Each target segment is also associated with corresponding spoken transcripts and a keyframe.

3.2 Linking Algorithm

The linking algorithm uses textual and visual features to determine the similarity between query and target anchors. We use metadata descriptions and ASR transcripts (LIUM and LIMSI) to represent textual information, and describe the visual content of keyframes using both low-level and high-level features.

Text Analysis We use the Apache Lucene 3.3.0[8] software in order to index and retrieve the segments based on textual information. ASR transcripts and

[4] http://www.vocapia.com/
[5] http://www-lium.univ-lemans.fr/en/content/language-and-speech-technology-lst
[6] http://www.multimediaeval.org/mediaeval2012/
[7] http://www.ffmpeg.org/
[8] http://lucene.apache.org/core/

metadata are merged into a single field for indexing of each segment. A standard analyzer of Apache Lucene is used to convert text data into the searching format. Text data in the single field is converted into lower case. The stop words are removed using the default list provided within Lucene. The analyzer tokenizes text based on a sophisticated grammar that recognizes e-mail addresses, acronyms, and alphanumerics [19]. The searching phase chops text data within of query anchor into terms and uses a *tf-idf* measure to score retrieved documents.

Low-level Visual Analysis We use a colour histogram and a bag-of-visual-word model to describe the low-level features of each keyframe. The colour histogram is calculated based on the HSV space. A three-level spatial pyramid representation is applied to each keyframe, which is divided into 1×1, 2×2, and 4×4 grids. The feature vector is normalized into $[0, 255]$, then a χ^2 function is applied to compare two histograms as following:

$$d(H_1, H_2) = \sum_{1 \leq i \leq k} \frac{(H_1(i) - H_2(i))^2}{H_1(i)} \tag{1}$$

where H_1 and H_2 represent two feature histogram respectively. The length of the feature vector is k, and $H_1(i)$ means the i^{th} point in histogram H_1.

The bag-of-visual-words model is generated by applying the SIFT descriptor [14] calculated by a total of 7,198 images randomly picked up from the video keyframe set. A K-means algorithm clusters the descriptor vectors to create visual words, where the number of cluster centres is experimentally set to 1,000. The weight vector of each keyframe is calculated based on visual words and its own SIFT descriptor. Finally, a cosine distance algorithm is applied to compute the distance between visual words.

High-level Visual Analysis We use two different high-level databases to extract the concepts (high level features) of each video keyframe. The first one is Object Bank[9] provided by Visual Lab, Stanford University. It contains a total of 177 high-level concepts created by a scale-invariant response map of a large number of pre-trained generic object detectors [13]. Each keyframe is described as a feature vector with the length of 44,604 which is calculated using multiple scales and different levels of a spatial pyramid. A Euclidean distance algorithm is applied to compute the distance between the high-level feature vectors.

The second high-level feature database is provided by the Vision Group at University of Oxford, specially created for the blip10000 dataset used in MediaEval 2012. It contains a set of concept detector scores for 589 concepts [4][10]. The detectors where trained by downloading positive images from Google images and learning their difference to assumed-to-be negative images in the dataset using the libLinear toolkit [8]. The distance between high-level concepts is calculated using the Euclidean distance.

[9] http://vision.stanford.edu/projects/objectbank/

[10] The concepts used were provided by Christoph Kofler from TU Delft.

4 Experimental Investigation

4.1 Crowdsourcing Task Design

Crowdsourcing allows us to obtain human-generated feedback about the relatedness between the video anchors, i.e. whether the hyperlinks that we create are valuable for real users. We collect feedback on whether users are interested in watching the selected video segment after having watched an initial query segment. Our investigation is carried out using the Amazon Mechanical Turk (AMT)[11] platform for crowdsourcing.

Traditionally, a task performed on the AMT platform is referred to as a Human Intelligence Task (HIT). In each HIT, our users were presented with a pair of video segments and were required to answer a number of questions to describe their opinion as to whether the two videos were related or not. Users were asked to provide details on the reason for their (un)relatedness judgement, and point out what features influenced their decision. We offered five options for the users to describe the feature selection: "Object", "Person", "Place", "Topic", and "Other" that can be the same in case of related videos or different in the case of unrelatedness. Moreover, in order to avoid spam submissions from workers and to determine reasonable answers from the workers, we also asked the workers to type in a number of meaningful words from the video segments that they had been asked to watch. The HIT reward was set at \$0.11, which was found to be acceptable to the workers.

4.2 Evaluation Overview

We uploaded a total of 8 runs to AMT for human evaluation involving different multimedia features, either textual or visual – as shown in Table 1. RUN_1, RUN_2, RUN_3 and RUN_4 use textual information to create video hyperlinks and RUN_5, RUN_6, RUN_7 and RUN_8 use low-level and high-level visual features.

A total of 3,915 HITs were created by all 8 runs. We received 3,521 useful submissions that were accepted for video hyperlinking evaluation. As working with videos is an unusual task on the AMT platform, we investigated the consistency of the decisions on video segment relatedness. This was based on the condition that each HIT was supposed to be answered by two different users. As it is possible to get a disagreement on the relatedness judgement, we defined that a pair of video segments is weakly related if only one user provides a positive answer on the relatedness judgement, whereas they are strongly related if the answers of both users are positive. There were 468 HITs marked as related. Within this set, 177 HITs were regarded as strongly related.

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

Table 1: Overview of the Video Hyperlinking Runs

RUN_NAME	Features	Types	
RUN_1	LIUM	Textual	
RUN_2	LIUM+META		
RUN_3	LIMSI		
RUN_4	LIMSI+META		
RUN_5	Colour Histogram	Low-level	Visual
RUN_6	Bag-of-Visual-Word		
RUN_7	Visual Group (Oxford)	High-level	
RUN_8	Object Bank (Standford)		

Table 2: User Options on the Relatedness Evaluation

OPTION	Object	Person	Place	Topic	Other
No. of Selection	243	244	247	430	133

Table 3: Overview of Positive Answers on Each Run. (WR: weak related, SR: strong related)

RUN	RUN_1	RUN_2	RUN_3	RUN_4	RUN_5	RUN_6	RUN_7	RUN_8
WR	64	67	60	65	44	53	41	29
SR	70	72	60	66	71	5	13	1
Total	134	139	120	131	115	58	54	30

Table 4: Overview of MAP Values. (WR: weak related, SR: strong related, ALL: WR+SR, WV: within the videos, WC: within the collection)

RUN		RUN_1	RUN_2	RUN_3	RUN_4	RUN_5	RUN_6	RUN_7	RUN_8
WV	ALL	**0.2108**	0.2084	0.1706	0.1919	0.1934	0.0562	0.0611	0.0329
	WR	**0.0597**	0.0564	0.0482	0.0559	0.0462	0.0469	0.0443	0.0324
	SR	**0.1107**	0.1072	0.0881	0.0940	0.1070	0.0039	0.0112	0.0006
WC	ALL	0.1209	**0.1293**	0.1082	0.1277	0.0753	0.0720	0.0692	0.0393
	WR	0.0496	0.0547	0.0468	0.0591	0.0302	**0.0622**	0.0553	0.0388
	SR	0.0406	**0.0416**	0.0362	0.0387	0.0266	0.0041	0.0080	0.0006

4.3 Evaluation Results and Analysis

Table 2 shows what features influence the relatedness judgement based on user feedback. The 'Object', 'Person' and 'Place' options mean that users determined the relatedness based on visual information, such as the same objects, location or human faces. The 'Topic' option means the users' judgement was influenced by the spoken information from video segments. Moreover, the 'Other' option was provided to allow users to express their own opinion on the relatedness judgement. Table 2 indicates that most users considered spoken information as an important aspect in evaluating hyperlinking relatedness.

Table 3 shows the number of relevant video segments retrieved by each run for both weak and strong relatedness. The query set used in the evaluation contains a total of 30 queries in each run. To evaluate the ranked list retrieved by each query, the top 10 video hyperlinking results were selected, with a total of 300

results for each run. According to table 3, the runs retrieved based on textual features achieved more positive results on the relatedness judgement. Among them, RUN_2 detects the most relevant video pairs, i.e. 139 out of 300 results. On the contrary, the performance of runs based on visual features decreases.

Average Precision (AveP) and Mean Average Precision (MAP) were used to evaluate the performance of each run. In addition to considering strong and weak relatedness, the evaluation also considers whether hyperlinks were created within the videos or within the collection. A hyperlink within a video means that a target segment exists either in the same video as the query anchor or in other different videos in the collection, while a hyperlink within the collection means a target anchor only exists in a different video. Table 4 shows an overview of MAP values for the different alternatives.

In general, the hyperlinking algorithms based on textual features perform better than those using visual features. The retrieval results using LIUM transcripts have the best score in most cases. An exception is the case of weak relatedness within the collection, where LIMSI with the corresponding video metadata achieves the best performance. MAP values and HIT feedback are consistent in the conclusion that speech data information is a bigger influence than visual data when judging the relevance of video segments. User feedback implies that they prefer to link two video segments that share the same or a similar story. The correspondence in person or object depicted is a much lower priority.

When comparing the results for visual features, both low-level feature descriptors, colour histogram and bag-of-visual words, always performs better than high-level feature descriptors. This is due to the fact that relevant video segments more easily share similar low-level visual features, such as background colour or illumination, while the performance of high-level features is seriously influenced by the Semantic Gap. This is clear when comparing the results for Visual Group (Oxford) and Object Bank (Stanford) high-level datasets. The former was specially created for the blip.tv dataset used in MediaEval 2012, whilst the latter, even if representative enough for a general image dataset, misses specific aspects within a TV dataset.

When we analyse results for within videos vs. within collection, there is a clear difference in terms of textual and visual features. Within videos, the best performance based on textual features is determined by the combination of ASR transcripts and metadata. Within the collection, LIUM+METADATA and LIMSI+METADATA show better performance than using single LIUM or LIMSI transcripts. When creating links within the same video due to the fact that metadata is always the same, the difference between spoken transcripts influences the ranked retrieval result. Within the whole collection, on the other hand, removing links locating in the same video, metadata information and ASR transcripts both exhibit differences in determining the description of video content. The performance of colour histogram features decreases significantly when linking videos within the whole collection. This is due to the fact that two video segments within the same video often share the same or similar background.

Table 5: Overview of AveP values

RUN	RUN_1	RUN_2	RUN_3	RUN_4	RUN_5	RUN_6	RUN_7	RUN_8
Topic_1	0.205	0.243	0.230	**0.252**	0.008	0.118	0.000	0.005
Topic_2	0.305	0.228	0.339	**0.385**	0.000	0.026	0.000	0.174
Topic_3	0.330	**0.356**	0.252	0.260	0.028	0.028	0.139	0.000
Topic_4	0.240	0.240	0.146	0.156	**0.455**	0.080	0.000	0.015
Topic_13	0.025	0.040	0.000	0.020	0.000	0.000	0.050	**0.400**
Topic_14	0.000	0.014	0.034	0.010	0.000	**0.347**	0.014	0.013

Table 6: Overview of Rerank LIUM, Colour Histogram, and High-level Concept MAP values, described as MAP/Increase Rate. (WR: weak related, SR: strong related, ALL: WR+SR, LM: LIUM transcripts+metadata, CH: colour histogram, VG: Visual Group (Oxford))

RUN	LM	LM+CH	LM+VG	LM+CH+VG
ALL	0.1293	0.1975 / +52.7%	0.1647 / +27.4%	0.2040 / +57.8%
WR	0.0547	0.1181 / +115.9%	0.0910 / +66.4%	0.1335 / +144.1%
SR	0.0416	0.0600 / +44.2%	0.0532 / +27.9%	0.0539 / +29.6%
RUN	CH	CH+LM	CH+VG	CH+LM+VG
ALL	0.0753	0.0927 / +23.1%	0.1312 / +42.6%	0.1265 / +68.0%
WR	0.0302	0.0577 / +47.7%	0.0628 / +107.9%	0.0644 / +113.2%
SR	0.0266	0.0170 / -36.1%	0.0434 / +63.2%	0.0360 / +35.5%
RUN	VG	VG+LM	VG+CH	VG+LM+OX
ALL	0.0692	0.0360 / -47.9%	0.0620 / -10.4%	0.0354 / -48.8%
WR	0.0553	0.0221 / -60.0%	0.0362 / -35.4%	0.0252 / -54.4%
SR	0.0080	0.0121 / +51.3%	0.0213 / +166.3%	0.0086 / +7.5%

Table 5 shows an overview of AveP values of each run for a total of 6 queries. All the AveP values are calculated for linking videos within the whole collection. Both weak relatedness and strong relatedness are considered. In general, AveP values are consistent with MAP evaluation. The retrieval results using the combination of LIUM/LIMSI transcripts and metadata information have better scores in the first three queries, while the scores decrease in the runs extracted by visual features. This demonstrates the conclusion that speech data has a higher priority when determining the relevance of a pair of video segments. However, in Topic_4, Topic_13, and Topic_14, the best performance is achieved by the runs using visual descriptors. In Topic_4, the run using colour histogram analysis has a score of 0.455. Figure 1 shows two groups of example keyframes associated with the retrieval results of Topic_4 in RUN_5 and Topic_14 in RUN_6. Figure 1(a) and Figure 1(b) present an introduction about certain software, with different spoken information but similar keyframes. Therefore, the analysis of visual content shows the advantage of removing the disagreement of voice messages and reflects the user's interest in the visual scene. Figure 1(c) and Figure 1(d) also suggest the same for visual high-level concepts. According to user feedback, the

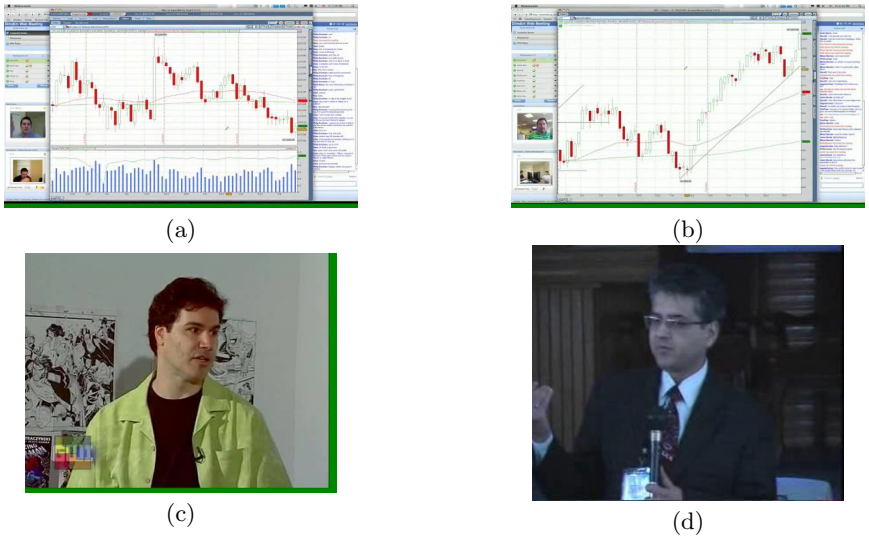

Fig. 1: Sample comparision of keyframes in Linked Video Segments

two video segments are regarded as relevant due to the fact that only one person gives a presentation, even if the content is quite different. Therefore, a further conclusion is that visual features can perform as a complement to textual feature analysis when constructing multimedia hyperlinks, and vice versa.

To prove this conclusion, a reranking algorithm was used to retrieve a new ranked list of the top 10 results implemented by different feature types. A total of 3 runs are selected based on LIUM ASR transcripts associated with corresponding metadata, colour histogram, and high-level concepts from Visual Group (Oxford). The top 10 results of each method were reranked by fusing normalized scores from the other two. A linear fusion algorithm was used where the weight for all scores was set to be equal. Table 6 shows the evaluation results based on the MAP measure. The reranking strategy improves most results comparing with Table 4. The improvement is clear for the linking strategies implemented by using LIUM transcripts and colour histogram. Based on these results we plan to carry out further work on multimedia hyperlinking by devising efficient fusion algorithms to utilize the advantage of textual and visual features.

5 Conclusions and Future Work

This paper describes our investigation into feature effectiveness for automatic multimedia hyperlinking. It simulates a scenario whereby the user browses a set of video data associated with existing hyperlinks across the whole collection. Our objective was to research how different multimedia features influence user performance and contribute to multimedia hyperlink generation. Automatic link

construction uses both textual and visual features, including LIUM/LIMSI transcripts, metadata information, colour histogram descriptor, bag-of-visual-words extracted by SIFT descriptor, and high-level visual concepts from the Visual Group (Oxford) and the Object Bank (Stanford). The evaluation is based on using human computing techniques supported by Amazon Mechanical Turk.

Crowdsourcing evaluation concludes that textual features exhibit better performance than visual features for multimedia hyperlink construction. The textual information related to a video can be extracted from both spoken data or metadata. In general, a combination of ASR transcripts and metadata shows the best results. Moreover, the quality of hyperlinks created based on visual features is variable. However, some potential links can be determined by visual features due to the lack of spoken information or incomplete metadata.

The evaluation suggests that textual information significantly contributes to the relevance of video segments. Searching and indexing spoken words should thus consider the context information and the concept of the story described by the whole video. Moreover, it is a challenge to efficiently fuse the results from different hyperlinking frameworks based on textual and visual features. Both of these aspects will form the basis of our future work.

6 Acknowledgements

This work was supported by funding from the European Commission's 7th Framework Programme (FP7) under AXES ICT-269980.and Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL) project at DCU.

References

1. O. Alonso, D. E. Rose, and B. Stewart. Crowdsourcing for Relevance Evaluation. *SIGIR Forum*, 42(2):9–15, Nov 2008.
2. M. Bron, B. Huurnink, and M. de Rijke. Linking Archives Using Document Enrichment and Term Selection. In *Research and Advanced Technology for Digital Libraries*, volume 6966, pages 360–371. 2011.
3. M. Bron, J. van Gorp, F. Nack, and M. de Rijke. Exploratory Search in an Audio-Visual Archive: Evaluating a Professional Search Tool for Non-Professional Users. In *1st European Workshop on Human-Computer Interaction and Information Retrieval (EuroHCIR 2011)*, 2011.
4. K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *Proceedings of the British Machine Vision Conference*, pages 76.1–76.12, 2011.
5. S. Chen, G. J. F. Jones, and N. E. O'Connor. DCU Linking Runs at Mediaeval 2012 Search and Hyperlinking Task. In *MediaEval*, volume 927 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2012.
6. M. Eskevich, G. J. F. Jones, R. Aly, R. J. Ordelman, S. Chen, D. Nadeem, C. Guinaudeau, G. Gravier, P. Sébillot, T. de Nies, P. Debevere, R. Van de Walle, P. Galuscakova, P. Pecina, and M. 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.

7. M. Eskevich, G. J. F. Jones, S. Chen, R. B. N. Aly, R. J. F. Ordelman, and M. Larson. Search and Hyperlinking Task at Mediaeval 2012. In *MediaEval 2012 Multimedia Benchmark Workshop, Pisa, Italy*, page 14, Aachen, 2012. CEUR-WS.org.
8. R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A Library for Large Linear Classification. *J. Mach. Learn. Res.*, 9:1871–1874, June 2008.
9. G. J. F. Jones. An Introduction to Crowdsourcing for Language and Multimedia Technology Research. In *Proceedings of the 2012 international conference on Information Retrieval Meets Information Visualization*, PROMISE'12, pages 132–154, Berlin, Heidelberg, 2013. Springer-Verlag.
10. P. Kelm, S. Schmiedeke, and T. Sikora. Feature-based Video Key Frame Extraction for Low Quality Video Sequences. In *10th Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2009, London, United Kingdom, May 6-8, 2009*, pages 25–28, 2009.
11. L. Lamel and J.-L. Gauvain. Speech Processing for Audio Indexing. In *Advances in Natural Language Processing*, pages 4–15. 2008.
12. M. Larson, E. Newman, and G. J. F. Jones. Overview of Videoclef 2009: New Perspectives on Speech-Based Multimedia Content Enrichment. In *Multilingual Information Access Evaluation II. Multimedia Experiments*, volume 6242, pages 354–368. 2010.
13. L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei. Object Bank: A High-Level Image Representation for Scene Classification and Semantic Feature Sparsification. In *Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2010.
14. D. Lowe. Object Recognition from Local Scale-Invariant Features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1150–1157, 1999.
15. R. Mihalcea and A. Csomai. Wikify!: Linking Documents to Encyclopedic Knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 233–242, 2007.
16. D. Milne and I. H. Witten. Learning to Link with Wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 509–518, 2008.
17. A. Rousseau, F. Bougares, P. Delglise, H. Schwenk, and Y. Estv. LIUM's Systems for the IWSLT 2011 Speech Translation Tasks. In *Proceedings of IWSLT 2011I*, 2011.
18. S. Schmiedeke, P. Xu, I. Ferrané, M. Eskevich, C. Kofler, M. A. Larson, Y. Estève, L. Lamel, G. J. F. Jones, and T. Sikora. Blip10000: A Social Video Dataset Containing SPUG Content for Tagging and Retrieval. In *Multimedia Systems Conference 2013, (MMSys '13)*, pages 96–101, 2013.
19. A. Sonawane. Using Apache Lucene to search text - Easily Build Search and Index Capabilities into your Applications. http://www.ibm.com/developerworks /library/os-apache-lucenesearch/, Aug 2009.
20. L.-X. Tang, I.-S. Kang, F. Kimura, Y.-H. Lee, A. Trotman, S. Geva, and Y. Xu. Overview of the NTCIR-10 Cross-Lingual Link Discovery Task. In *Proceedings of NTCIR-10*, 2012.