Which Information Sources are More Effective and Reliable in Video Search

Zhiyong Cheng[‡], Xuanchong Li[§], Jialie Shen[‡], Alexander G. Hauptmann[§]
[‡]School of Information Systems, Singapore Management University, Singapore
[§]School of Computer Science, Carnegie Mellon University, USA
{jason.zy.cheng, jialie}@gmail.com, {xcli, alex}@cs.cmu.edu

ABSTRACT

It is common that users are interested in finding video segments, which contain further information about the video contents in a segment of interest. To facilitate users to find and browse related video contents, video hyperlinking aims at constructing links among video segments with relevant information in a large video collection. In this study, we explore the effectiveness of various video features on the performance of video hyperlinking, including subtitle, metadata, content features (i.e., audio and visual), surrounding context, as well as the combinations of those features. Besides, we also test different search strategies over different types of queries, which are categorized according to their video contents. Comprehensive experimental studies have been conducted on the dataset of TRECVID 2015 video hyperlinking task. Results show that (1) text features play a crucial role in search performance, and the combination of audio and visual features cannot provide improvements; (2) the consideration of contexts cannot obtain better results; and (3) due to the lack of training examples, machine learning techniques cannot improve the performance.

Keywords

Video Search, Video Hyperlinking

1. INTRODUCTION

With the explosive growth and the widespread accessibility of multimedia contents on the Web, video is becoming one of the most valuable sources to assess information and knowledge [4, 19]. When consuming video content, users are highly interested in finding further information on some aspects of the topics of interest associated with a video segment. Therefore, it is crucial to develop effective video search and hyperlinking to help users explore, navigate and search interesting video contents in audiovisual archives. *Video hyperlinking* aims at linking a video segment (or video anchor) to other video segments in a video collection, based on similarity or relatedness. Formally, the video hyperlinking task is defined as[1]: *given a set of test videos with metadata and a defined set of anchors, each defined by start time and end time in the video, return for each anchor a ranked list of hyperlinking targets.* Therefore, video hyperlinking enables users to navigate between video segments in a large video collection [3].

To facilitate the development and advancement of video hyperlinking systems, video hyperlinking has become a competition task since 2012 in MediaEval [6]. Standard test collections are provided and metrics are defined for the evaluation of developed systems. The task is defined to find *relevant anchors* or *short segments* (e.g., 2 minutes) of video contents given a set of query anchors. Thus, the hyperlinking is generally addressed within an information retrieval framework. As videos in the test collection could be in hours of length, video hyperlinking consists of two steps: (1) *video segmentation* - separate a video into a number of short video segments and (2) *video retrieval* - retrieval potential links to videos or video segments.[2] For video segmentation, many systems apply fixed-length segmentation to separate videos into fixed-length of segments. Other video segmentation methods have also been developed and studied, such as *video shot based* and *semantic-based segmentation*, however, no evidence shows that those segmentation methods are better than the fix-length segmentation method. More research efforts were devoted into the development of effective retrieval methods, including the explorations of different information sources (e.g., subtitle, metadata, transcriptions, segment surrounding context, name entity, enrichment of concept and synonyms, as well as audio and visual features [8]) and search strategies (e.g., combination with or re-ranking with visual features, combination of video-level and segment-level retrieval).

In this paper, we use the fixed-length video segmentation method and mainly focus on studying the effects of different types of information sources on the performance of video hyperlinking, including text (subtitle, metadata, transcription) and a variety of video content (audio, visual and motion) features. Nine different text-based retrieval methods are used based on the text information with and without the consideration of surrounding context (around the query or target segment). Besides, we also study the performance of multimodal feature combination using weighted linear combina-

SIGIR '16, July 17–21, 2016, Pisa, Italy.

DOI: http://dx.doi.org/10.1145/2911451.2914765

[1]The definition is from TRECVID 2015 Video Hyperlinking task, http://www-nlpir.nist.gov/projects/tv2015/#lnk.

[2]The order of the two steps can be the reverse, retrieving relevant videos firstly and then extracting the most relevant segments from those videos identified in the first step [5].

tion. Further, we evaluate the performance of different combination weights over different categories of queries, which are classified according to their contents. Experiments show that surrounding context and video-content features have little contribution on the performance improvement.

2. VIDEO HYPERLINKING SYSTEM

In this study, we address the video hyperlinking as an ad-hoc retrieval problem. Given a query anchor indexed with certain features, video segments in the test collection are also indexed with the same features, and then a retrieval method is used to search and return the most relevant video segments to this query. In our experiments, we (1) first separate each video in the collections into *50s fixed-length* segments without overlapping, as this configure achieved good performance in the CUNI2014 video hyperlinking system [8]; (2) different types of features are extracted from each segment; (3) a variety of retrieval methods are explored; and (4) different strategies are used to combine the results obtained based on different features. In the next, we describe the features used and the retrieval methods considered in experiments.

2.1 Dataset

We first introduce the dataset used in the study. Our study is performed on the dataset of TRECVID 2015 Video Hyperlinking task. For the ease of understanding, we first clarify several terminologies: (1) *video* refers a long video clip (usually longer than 20 minutes); (2) both *video segment* (or *segment*) and *video anchor* (or *anchor*) refer to a short segment of a video (usually less than 2 minutes), defined by a start timestamp and an end timestamp; (3) *video metadata* (or *metadata*) contains the *title* and *a short program description* of a video; (4) *subtitle* is the manual annotation of the speech in a video; and (5) *transcript* refers to the annotation obtained by using automatic speech recognition (ASR) methods.

The TRECVID 2015 Video Hyperlinking dataset consists of around 2686 hours of BBC television broadcasts content from 05/12/2008 to 07/31/2008. The data is accompanied by metadata, subtitle, three kinds of ASR transcripts (generated by LIMSI [12], LIUM [16], and NST-Sheffield [14] respectively), two versions of key concepts (detected by two concept detectors), and the prosodic audio features [7].

The development set contains 30 query anchors. For each of them, a set of ground-truth anchors is provided. The number of positive segments for each query anchor varies from 17 to 122. Notice that many positive segments are from the same video from which the corresponding query anchor is extracted. The test set contains 135 query anchors, which are for the final evaluation in the competition.

2.2 Retrieval Methods

2.2.1 Text-based Methods

Text Features We explore the effectiveness of different sources of textual features, namely the subtitle and three kinds of ASR transcripts. For each type of the features, we also consider their combinations with metadata and surrounding contexts. The tested lengths of surrounding context include 50s, 100s, and 200s. Hence, for each of subtitle, LIMSI, LIUM, and NST-Sheffield, there are eight indexing methods.[3] For a video segment, the combination of *subtitle and metadata* is to concatenate *the subtitle of this segment* with *the metadata of the video* from which the segment is extracted. Similarly, the combination of *subtitle, metadata*, and *50s surrounding context* is the concatenation of *the subtitle of this segment*, *50-seconds-length passage before and after the segment*, and *the metadata of the corresponding video*. All the textual sources are processed by punctuation & stop-words removal and capitalization normalization.

Retrieval Methods For each type of feature, we experiment with nine different retrieval models: (1) BM25, (2) DFR version of BM25(DFR-BM25) [9], (3) DLH hypergeometric DFR model (DLH13) [1], (4) DPH [2], (5) Hiemastra's Language Model (Hiemastra-LM) [11], (6) InL2 - inverse document frequency model for randomness, Laplace succession for first normalization, and normalization 2 for term frequency normalization [9], (7) TF-IDF, (8) LemurTF-IDF[4], and (9) PL2 - poisson estimation for randomness, Laplace succession for first normalization, and normalization 2 for term frequency normalization [9]. We used Terrier[5] IR system to run experiments with these retrieval methods (with default parameters) with different textual sources.

2.2.2 Content-based Method

For the content-based method, we use various video features in the retrieval task. These video features include motion features (e.g., improved dense trajectory [13]), audio features (e.g., MFCC) and visual semantic features [15]. After explicit feature mapping [18], the cosine similarity is used as the relevance score.

2.2.3 Multimodal-based Method

We explore the effects of the combination of different features, based on the assumption that different features can capture different aspects of video content.

Weighted Linear Combination (WLC) The relevant score of a video segment with respect to a query is computed by a weighted linear combination of the relevant scores obtained by different features. Let $wlc(q, v)$ be the final relevance score obtained by the weighted linear combination, and $rel(f_i)$ be the relevance score obtained based on feature f_i. Given the selected feature $\{f_1, f_2, \cdots, f_n\}$, the $wlc(q, v)$ is computed by:

$$wlc(q, v) = w_1 \cdot rel(f_1) + w_2 \cdot rel(f_2) + \cdots + w_n \cdot rel(f_n) \quad (1)$$

where $\{w_1, w_2, \cdots, w_n\}$ are the linear combination weights, which characterize the contribution of different features on the final performance. These weights are tuned on the training set. Due to the lack of training examples (refer to Table 1), we only used five features in our experiments. These features are selected based on on their individual performances and the consideration of feature heterogeneity. Specifically, the selected features are: *subtitle_metadata_LemurTF-IDF*[6], *subtitle_metadata_DPH*, *keyconcept_Lemur*

[3] Taken *subtitle* as an example, there are *subtitle*, *subtitle with 50s context*, *subtitle with 100s context*, *subtitle with 200s context*, *subtitle and metadata*, *subtitle and metadata with 50s context*, *subtitle and metadata with 100s context*, and *subtitle and metadata with 200s context*.

[4] http://www.lemurproject.org/

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

[6] Subtitle_metadata_LemurTF-IDF denotes that the relevant score is obtained by LemurTF-IDF based on the subtitle and metatada. Similar definition is applied for other methods.

TF-IDF, *improved trajectory*, and *MFCC*. Keyconcept_Lemur TF-IDF denotes the TF-IDF method based on the key concepts of keyframes. The *key concepts* are the concepts detected in the keyframes with normalized scores greater than 0.7, using the Leuven's concept detectors of 1537 ImageNet concepts [17]. For a video segment, its key concept based representation is the concatenation of key concepts detected in all the keyframes of this segment.

For different types of videos, their contents or topics could be very diverse. Different types of features contribute to the representation of video contents differently. It would be useful to use different weights for different video categories. Thus we classify the videos into categories based on the programme category ontology of BBC news. Due to the limited query examples in the development dataset, we further group the videos into two broad categories:

- Category 1: news & weather; science & nature; music (religion & ethics); travel; politics news; life stories music; sport (tennis); food & drink; motosport.
- Category 2: history; arts, culture & the media; comedy (sitcoms), cars & motors; antiques, homes & garden, pets & animals; health & wellbeing, beauty & style.

In general, videos in the sub-categories of Category 1 enjoy more similar contents in text, audio, and visual features (such as news and music), and thus queries in Category 1 are easier to achieve better results. In contrast, for videos in the same sub-categories of Category 2, although their contents are about the same topic, the contents could be diverse in contents. For example, videos about *history* or *health* could be very different in words and scenes.

To evaluate the performance of this method, we randomly split the query anchors in the development set into training set and testing set. Table 1 presents the details of training set and testing set for global weighted linearly combination (GWLC - without the consideration of video categories) and categorized weighted linear combination (CWLC). Notice that the number of training examples is very small, especially for the CWLC method, which limits the performance of the weighted linear combination.

Table 1: Sizes of training set and testing set for the GWLC (whole) and CWLC (category 1 and category 2) methods.

Category	# Queries in the Training Set	# Queries in the Testing Set
GWLC	15	15
Category 1	9	9
Category 2	8	4

Learning to Rank [10] We also explored the use of learning to ranking techniques for refining the retrieval results. The retrieval scores obtained by different features are used as the input of different learning algorithms (e.g., linear regression, Naive Bayes, SVM, etc.). Unfortunately, these techniques cannot improve the performance, due to the lack of well-labeled data. Due to the space limitation, we have not reported the results of these methods in the following.

3. EXPERIMENTS

To evaluate the performance of video hyperlinking systems, the top ranking results of submissions are accessed using a mechanical turk (MT) crowdsourcing approach. A test assessment on a smaller part of the data by a local team of target users is used to identify potential discrepancies between the MT workers' judgments and those of the target user group. Descriptions given by the anchor creators (anchor descriptions, description and format requested targets) are used for evaluation purpose. In the generation of ground truth, only a subset of the submissions for each query will be used in evaluation. To reduce the workload of evaluators, for the anchors longer than 2 minutes, only the first two minutes will be used as the basis of relevance assessment. For more details about hyperlinking evaluation, please refer to [5]. The submissions are evaluated based on the precision at a certain rank measure, adapted to unconstrained time segments. In this paper, we report the performance on evaluation metrics of Precision@$\{5, 10, 20\}$ and MAP.

In the next, we report the experiment results of different methods on dataset. The results of content-based methods have not presented because of the overall poor performance.

Table 2: Results of using different transcripts, metadata, retrieval methods and contexts. In each row, the retrieval method is the best retrieval methods among the nine tested methods for the corresponding text source. 50s, 100s and 200s refer to the lengths of contexts. Please refer to Sect. 2.2.1 for the retrieval method in the "Method" column: (1) BM25, (3) DLH13, (4) DHP, (5) Hiemastra-LM, (8) LemurTF-IDF, and (9) PL2. NST refers to NST-Sheffield transcript.

Text Information	Method	MAP	P@5	P@10	P@20
Subtitle	(8)	**.162**	**.324**	**.297**	**.228**
LIMSI	(8)	.093	.215	.173	.137
LIUM	(1)	.056	.144	.124	.098
NST	(8)	.065	.164	.129	.102
Subtitle_Metadata	(8)	**.197**	**.293**	**.253**	**.205**
LIMSI_Metadata	(8)	.147	.200	.173	.147
LIUM_Metadata	(4)	.107	.147	.157	.132
NST_Metadata	(8)	.123	.153	.147	.128
Subtitle_50s	(9)	.114	.173	.137	.118
Subtitle_100s	(5)	.124	.220	.170	.132
Subtitle_200s	(3)	.128	.227	.160	.103
Subtitle_Metadata_50s	(3)	.124	.200	.147	.112
Subtitle_Metadata_100s	(5)	**.136**	.220	.180	**.135**
Subtitle_Metadata_200s	(3)	.134	**.247**	**.194**	.113

3.1 Performance on Development Data

Text-based Retrieval Method Table 2 shows the results of text-based retrieval methods using different text sources. For each type of text source, only the best performance obtained by the nine retrieval methods is reported. As a large set of text-based retrieval methods (different text sources and different retrieval methods) has been explored, we have not presented the results of all methods. The results are grouped into three groups in the table. As the performance of using subtitle is much better than the use of ASR transcripts (LIMSI, LIUM and NST-Sheffield), we did not show the performance of ASR transcripts with the consideration of context. The performance based on ASR transcripts is limited by the speech recognition accuracy. Among the three ASR transcripts, LIMSI obtains the best performance, followed by LIUM. Not surprising, with the consideration of metadata, the performance of ASR transcripts can be significant improved, as the metadata is manually annotated and summarizes the video contents. The combination of subtitle and metadata obtains higher MAP

over subtitle while slightly lower precisions on top results. Recall that the metadata is a summary of the contents in a video. Supppose a query segment v_q is extracted from video V and a video segment v_s is extracted from V'; and V and V' is about the same topic, namely, their metadatas contain very similar content. The consideration of metadata in retrieval will increase the similarity score of v_s with respect to the query v_q, and thus move the video segment to a higher ranking position in the result list. If v_s is irrelevant, the consideration of metadata may cause the video segment v_s in a relatively high position, resulted in the decrease of precision; on the other hand, if v_s is indeed relevant to the query v_q, the resulted higher position due to metadata will lead to the increase of MAP.

From the results of the third group in the table, the consideration of context data makes the performance significantly decreased. The results imply that the incorporation of context data introduces noisy data, which misleads the search of relevant segment. By comparing the search methods of different text sources, it can be found that better performances are obtained by the vector space (LemurTF-IDF) method for text information without context (i.e., relatively short documents), and better performances are obtained by probabilistic methods with the consideration of contexts (i.e., relatively long documents).

Weighted Linear Combination Table 3 reports the performance of weighted linear combination methods. Because the performances of different queries varied in large ranges, we list the corresponding performance of the test queries using Subtitle_Metadata_LemurTF-IDF for comparisons. Obviously, the queries from Category 1 obtained much better results than queries from Category 2. By comparing with the performance of weighted linear combination methods, we can observe that the performance decreases with the combination of other features based on the simple late fusion method.

Table 3: Results of weighted linear combination methods.

Method	MAP	P@5	P@10	P@20
LemurTF-IDF	.305	.369	.339	.281
GWLC	.270	.400	.377	.327
LemurTF-IDF (category 1)	.432	.467	.456	.383
CWLC (category 1)	.381	.511	.489	.444
LemurTF-IDF (category 2)	.020	.150	.075	.050
CWLC (category 2)	.020	.150	.100	.0625

3.2 Performance on Test Data

In this section, we present the results of three methods on the test data in the final evaluation of the TRECVID 2015 Video Hyperlinking task: (1) Subtitle_Metadata_LemurTF-IDF (SM_LemurTF-IDF), (2) Global Weighted Linear Combination (GWLC), (3) Categorized Weighted Linear Combination (CWLC). Table 4 shows the results of the submitted runs on the test data. The same conclusions can be observed: the best performance is obtained by only using textual data. It is worth mentioning that the results of these methods achieved top positions in the competition.

Table 4: Performance on the test data.

Method	MAP	P@5	P@10	P@20
SM_LemurTF-IDF	.462	.654	.608	.438
GWLC	.316	.630	.534	.403
CWLC	.313	.630	.524	.401

4. CONCLUSION

In this paper, a large set of textual and video content features on the performance of video hyperlinking has been studied. The results show that the video hyperlinking performance relies on manual annotations (subtitle and metadata). The performance based on the ASR transcriptions is still far from the one achieved by manual annotations, while it is much better than audio, visual and motion features. The combination of surrounding context information will decrease the performance. The use of video content based features (audio, visual, and motion) has negative effects on the performance of textual features. Due to the limitation of well-labeled data, it is difficult to study the effectiveness of machine learning techniques on video hyperlinking task.

5. REFERENCES

[1] G. Amati. Frequentist and bayesian approach to information retrieval. In *Proc. of ECIR*, 2006.

[2] G. Amati, G. Amodeo, M. Bianchi, Ca. Gaibisso, and G. Gambosi. Fub, iasi-cnr and university of tor vergata at trec 2008 blog track. In *Proc. of TREC 2008*, 2008.

[3] C. Bhatt, N. Pappas, M. Habibi, and A. Popescu-Belis. Multimodal reranking of content-based recommendations for hyperlinking video snippets. In *Proc. of ACM ICMR*, 2014.

[4] X. Chang, Y. Yu, Y. Yang, and A. Hauptmann. Searching persuasively: Joint event detection and evidence recounting with limited supervision. In *Proc. of ACM MM*, 2015.

[5] M. Eskevich, G. Jones, and R. Aly. Multimedia information seeking through search and hyperlinking. In *Proc. of ACM ICMR*, 2013.

[6] M. Eskevich, G. Jones, S. Chen, R. Aly, R. Ordelman, and M. Larson. Search and hyperlinking task at mediaeval 2012. In *Proceedings of MediaEval Workshop*, 2013.

[7] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proc. of ACM MM*, 2013.

[8] P. Galuščáková, P. Pecina, M. Kruliš, and J. Lokoc. Cuni at mediaeval 2014 search and hyperlinking task: visual and prosodic features in hyperlinking. In *Proceedings of the MediaEval Workshop*, 2014.

[9] A. Gianni. Probabilistic models for information retrieval based on divergence from randomness. *PhD Thesis, School of Computing Science, University of Glasgow*, 2003.

[10] L. Hang. A short introduction to learning to rank. *IEICE TRANSACTIONS on Information and Systems*, 94(10):1854–1862, 2011.

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

[12] L. Lamel. Multilingual speech processing activities in quaero: Application to multimedia search in unstructured data. In *Proc. of Baltic HLT*, pages 1–8, 2012.

[13] Z. Lan, M. Lin, X. Li, A. Hauptmann, and B. Raj. Beyond gaussian pyramid: Multi-skip feature stacking for action recognition. *arXiv preprint arXiv:1411.6660*, 2014.

[14] P. Lanchantin et al. Automatic transcription of multi-genre media archives. In *First Workshop on Speech, Language and Audio in Multimedia*, 2013.

[15] P. Over, J. Fiscus, G. Sanders, D. Joy, M. Michel, G. Awad, A. Smeaton, W. Kraaij, and G. Quénot. Trecvid 2014–an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID*, 2014.

[16] A. Rousseau, P. Deléglise, and Y. Estève. Enhancing the ted-lium corpus with selected data for language modeling and more ted talks. In *The 9th edition of the Language Resources and Evaluation Conference*, pages 3935–3939, 2014.

[17] T. Tommasi et al. Beyond metadata: searching your archive based on its audio-visual content. In *Proceedings of the 2014 International Broadcasting Convention*, 2014.

[18] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(3):480–492, 2012.

[19] S. Yu, L. Jiang, Z. Xu, Y. Yang, and A. Hauptmann. Content-based video search over 1 million videos with 1 core in 1 second. In *Proc. of ACM ICMR*, 2015.