Exploring Video Hyperlinking in Broadcast Media

Maria Eskevich, Quoc-Minh Bui, Hoang-An Le, Benoit Huet
EURECOM, Sophia Antipolis, France
huet@eurecom.fr

ABSTRACT

Multimedia content produced on a daily basis and in constantly growing quantity by professionals and individual users, requires creation of navigation systems that allow access to this data on different levels of granularity in order to contribute to further discovery of a topic of interest for the user or to facilitate individual user browsing within a collection.

In this paper we describe our approach to enable users to browse through the multimedia collection. We implement the hyperlinking approach that uses the fine-grained segmentation of the visual content based on the scene segmentation, as well as available metadata, transcripts, and information about extracted visual concepts.

The approach was tested at the MediaEval Search and Hyperlinking 2014 evaluation task, where it has shown its effectiveness at locating accurately relevant content in a large media archive.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems; H.5.4 [**Information Interfaces and Presentation**]: HyperText/Hypermedia

Keywords

Multimedia Search; Media Fragment; Hyperlinking

1. INTRODUCTION

Multimedia content steadily increases its share in the overall Internet traffic [1]. This is due to the growing number of available online media platforms featuring professional (i.e. broadcast) and user contributed content, together with the amount of media content available for consumption. These platforms and high speed Internet connection allow users to watch comfortably a relevant video, once it is found in the collection. However, the actual retrieval of the content that would be relevant to the user's request, and that could allow each individual user to follow their own path of interest within the hyperlinked collection, still remains a challenging task, as it is less straightforward due to the lack of interpretability of the content.

Most often, videos within collections are described as a whole by textual information, hence pointing out at two weaknesses of actual systems: first, due to the semantic gap, the retrieval process is based on text features, while the rich multimodality of videos is not exploited. Second, it is not possible to partially retrieve a video, i.e., to obtain a specific fragment without having to watch the entire video. Therefore we follow our approach that used scene-based segmentation in combination with metadata and visual concepts information that allows operate at finer-temporal scale and takes video content into account throughout the process [2].

In our experiments we use the test dataset from the MediaEval 2014 Search and Hyperlinking task [7], containing 3 528 broadcast programs from various genre and totaling 2 686 hours of content that is processed by our system as presented in Section 3. These media files are indexed and processed as a collection of topically coherent media fragments in order to become a network of entirely interconnected anchors (i.e. currently played media fragments) and hyperlinks (i.e. potential video fragments to jump to in order to get more information on the anchor). This approach is tested by creating hyperlinks for officially provided 30 anchors of the benchmarks.

The remainder of the paper is organized as follows: Section 2 provides the context of the task and potential solutions available in the field, Section 3 describes the system architecture with the media analysis components, Section 4 introduces the experimental results, that are discussed in Section 5, and Section 6 concludes our findings.

2. RELATED WORK

Search and hyperlinking tasks within a video collection have been increasingly researched over the past years, in particular at the instigation of the MediaEval benchmark[2]. Different techniques have been studied to perform those tasks, the differences mainly being the video segmentation method and the features used for retrieval. As most of this data processing is currently done offline over the collection, and same systems are developed to deal with both search and hyperlinking, these two tasks are considered akin differing mostly at the stage of querying the system and further interaction by the user with the system output [19].

[1] Cisco Visual Networking Index

[2] www.multimediaeval.org

SLAM'15, October 30, 2015, Brisbane, Australia.

ACM 978-1-4503-3749-6/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2802558.2814647.

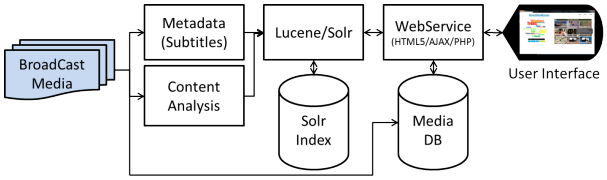

Figure 1: System Architecture Overview

Basic segmentation of video using the automatic or manually created transcript is the most employed technique, it comprises fixed-length segments with sliding windows, sentences or speech segments, etc [30, 10, 17, 9, 23, 22]. Diverse variations of these methods have been studied and cross compared: in [9], the authors intend to adjust starting points of the segments in order to match full sentences, using speech segment boundaries and pauses. The work of Schouten et al. [23] goes in the same direction: they build a probabilistic framework to model the importance of words and refine segment boundaries accordingly.

Nevertheless, some algorithms intend to segment videos into meaningful segments, based on topics derived from transcripts [10, 12, 3]: in [10, 11], the authors use classification trees to define the starting and ending times of these segments; [12, 26] exploit lexical cohesion within segments. On a similar fashion, our work includes a scene segmentation technique, although it differs by the fact that it is based on *visual* and temporal coherence of the video segments that constitute the scenes of the video.

Regarding the features used for retrieval, most studies have based on text similarity (vector-based models and TF-IDF weightings). An interesting direction taken by some works is to enrich the initial text by making use of diverse annotation methods, such as named entities or synonyms [5, 17, 4]. They rely on an offline step of pre-processing and document annotation, before performing queries.

Last, visual information is shown to give very small improvement in the hyperlinking process [17, 29, 3, 29]. Possible features used are visual concepts or SURF and SIFT features, detected at the shot level. In [3], hyperlinking results derived from text algorithms are re-ranked based on visual similarity with the anchor query. Similarly to those works, we intend to use visual features to improve the ranking of results of the hyperlinking, but our approach offers novelty in including visual analysis during the search process.

3. SYSTEM ARCHITECTURE

The architecture of the our system, as depicted in figure 1, is composed of both offline and online processing components. Multimodal content analysis and indexing (using Lucene/Solr[3]) is performed offline, whenever a new video is added to the archive, while the Web-service issues queries to the Solr index at run-time corresponding to the user activity.

3.1 Offline pre-processing and indexing

First we worked on dataset pre-processing. We applied techniques for scenes segmentation, concepts detection, keywords extraction, optical character recognition (OCR), face detection and tracking and named entities recognition. The idea was to extract as much information as possible, aiming to have a huge pool of features as input to the retrieval algorithms. We indexed all outcomes of this processing in a Lucene index[4] to store available information in a unified way. Such a methodology enables to easily test different algorithms for retrieval, thus making it possible to focus on the design of the algorithm. Nevertheless, not all information has been used in the methods reported in this paper. In this section, we only describe the techniques that were used in our framework.

3.2 Content Analysis

When ingesting a new multimedia document into our system, it is stored into the media database and processed at two levels: the entire document and the media fragment. At the level of the entire document, various types of metadata are available, i.e. title, cast, description, broadcast time, etc. The media fragments are defined using a complex segmentation procedure.

3.2.1 Scene Segmentation

Having started from a decomposition of each video of the collection into shots (using the provided automatic shot segmentation results [6]), and we aimed to define a more meaningful decomposition of each video into story-telling parts. For this we used the scene segmentation algorithm of [25]. This method groups shots into sets that correspond to individual scenes of the video, based on the content similarity and the temporal consistency among shots. Shot similarity in our experiments meant visual similarity, and the latter was assessed by computing and comparing the HSV histograms of the keyframes of different shots. Visual similarity and temporal consistency are jointly considered during the grouping of the shots into scenes, with the help of two extensions of the Scene Transition Graph (STG) algorithm [31]. The first extension, Fast STG, reduces the computational cost of STG-based shot grouping by considering shot linking transitivity and the fact that scenes are by definition convex sets of shots. The second one, Generalized STG, builds on the former to construct a probabilistic framework that alleviates the need for manual STG parameter selection, while also making possible the combination of different, heterogeneous approaches to evaluating shot similarity (i.e. using not only low-level visual features, as we did in our experiments, but also visual concepts, low-level audio features and audio events, for example).

3.2.2 Concept Detection

For visual concept detection, we follow the approach proposed in [24], using a sub-set of 10 base detectors per concept and a set of 151 semantic concepts, both static and dynamic ones, selected from the list of concepts defined in the TRECVID 2012 SIN task [20]. The 10 base detectors are applied at keyframe level (one keyframe per shot) and exploit different features (combinations of interest point detectors [13], descriptors such as SIFT [18], RGB-SIFT and Opponent-SIFT [27], and visual word assignment methods [28]). When evaluating a non-annotated shot, each base detector for a given concept calculates a Degree of Confidence (DoC) score in the range [0, 1], expressing the classifier's confidence in this concept being suitable for annotating

[3] `http://lucene.apache.org`

[4] `http://lucene.apache.org/solr/`

Table 1: Mean Precision @ 5 and @ 10 ranks for different relevance types (overlap, binned, tolerance)

Method ID	Mean Precision @ 5			Mean Precision @ 10		
	Relevance type					
	overlap	binned	tolerance	overlap	binned	tolerance
MM_VS_M	0.3000	0.2923	0.2538	0.2825	0.2769	0.2500
MM_TS_M	0.3538	0.3615	0.3385	0.2654	0.2692	0.2385
Text_VS_M	**0.5040**	**0.4800**	**0.4160**	**0.4480**	**0.4080**	**0.3600**
Text_S_MLT1_I	0.3000	0.3308	0.2615	0.2462	0.2615	0.2231
Text_S_MLT1_M	0.4167	0.4083	0.2917	0.3750	0.3625	0.2750
Text_S_MLT1_S	0.3000	0.3071	0.2571	0.2857	0.2857	0.2321
Text_S_MLT1_U	0.2692	0.2846	0.2385	0.2577	0.2731	0.2192
Text_S_MLT2_I	0.2333	0.2600	0.2133	0.1833	0.1967	0.1500
Text_S_MLT2_M	0.3667	0.3733	0.3000	0.3267	0.3167	0.2600
Text_S_MLT2_S	0.2067	0.2067	0.1600	0.2233	0.2267	0.1800

the current shot. The 10 computed DoC scores are then averaged to generate the final detection score for the concept. This process is iterated for all (151) considered concepts, and the vector of 151 final detection scores is the output of the concept detection component.

3.3 Hyperlinking

Hyperlinking is accomplished by automatically crafting a multimodal query from the currently played media fragment. The text query is compiled by extracting keywords from the subtitles aligned between the start and end time of the media fragment. Visual concepts scores that are taken from the corresponding indexed data of the key-frames containing in the media fragment. If the media fragment contains more than one shot, the highest score over all shots for each concept is used.

4. EXPERIMENTAL RESULTS

Two main types of the approaches that combine different variations of data processing are marked as following: 1) MoreLikeThis (MLT) Solr extension, and 2) using SolrâĂŹs query engine. MLT is used in combination with the sentence segments (S), using either text (MLT1) or text and annotations (MLT2). When Solr is used directly, we consider text only (Text) or with visual concept scores of anchors (MM) to formulate queries into the system. Keywords appearing within the anchor subtitles compose the textual part of the query. Visual concepts whose scores within the query anchor exceed the 0.7 threshold are identified as relevant to the video anchor and added to the Solr query. Both visual (VS) and topic scenes (TS) granularities are evaluated in this approach. When searching through the collection we use different provided transcripts (I: LIMSI/Vocapia [14], U: LIUM [21], S: NST-Sheffield [15], M: manually produced subtitles).

5. DISCUSSION

Table 1 shows the results across different runs in terms of mean precision at ranks 5 and 10, when different relevance techniques were used [1].

The best results were scored across all the metrics for the approach "Text_VS_M' that used the visual scenes and the manual subtitles for the hyperlinkins retrieval.

When MLT approaches are compared, we can see that the MLT1 outperforms MLT2 for the same type of transcript being used. This decrease in performance can be explained by the fact that the annotations added in MLT2 are mostly video based, and our work on the granulated media fragments requires annotation to be extracted or created with more precise time allocation.

Even though the runs using the visual content processing do not reach the highest scores, we can see their potential, as they reach comparable scores to the purely text based runs. Interestingly, the combination of visual concept scores of anchors with topic scenes improves over the visual scenes results when only the top 5 ranks are taken into account, and the trend becomes the opposite when the top 10 ranks are considered.

Currently we use a set of visual concepts defined in the TRECVID 2012 SIN task. These can be further elaborated when trained using the deep learning approach, and expanded through the use of transcript content partially describing or defining the visual content in the videos.

Overall, as the results of our algorithm are ranked lists of video segments presented to the users, it is worthwhile to question the impact of the segmentation used. The approaches using scenes have shown higher performances when compared with the more granular though too specific shot segmentation or higher level of videos [8, 16]. Scenes have been segmented by taking into account temporal and visual coherence, hence are suitable for meaningful fragments proposition, as the users should appreciate the smooth development of a story when following the path of hyperlinks chosen by him, and not a disrupted collection of video fragments.

6. CONCLUSION

While popular search engines retrieve documents on the basis of text information only, this paper aimed at proposing and evaluating an approach to include visual properties in the search of video segments. Experimental results, conducted on the MediaEval 2014 Search tasks, show that mapping text-based queries to visual concepts is not a straightforward task. Manually selecting relevant concepts is required impractical human intervention and does not necessarily lead to perfect results. The system performed among the best of MediaEval Search and Hyperlinking task, indicating the relevance and accuracy of our hyperlinking method. The proposed framework operates on the user query at the keyword level. As future work, we intend to further incorporate the semantic of the query when identifying key visual

semantic concepts based on named entity recognition approaches.

7. ACKNOWLEDGEMENTS

This work has been partially supported by Bpifrance within the NexGen-TV Project, under grant number F1504054U.

8. REFERENCES

[1] R. Aly, M. Eskevich, R. Ordelman, and G. J. F. Jones. Adapting binary information retrieval evaluation metrics for segment-based retrieval tasks. *CoRR*, abs/1312.1913, 2013.

[2] E. Apostolidis, V. Mezaris, M. Sahuguet, B. Huet, B. Cervenková, D. Stein, S. Eickeler, J. L. Redondo Garcia, R. Troncy, and L. Pikora. Automatic fine-grained hyperlinking of videos within a closed collection using scene segmentation. In *ACMMM 2014, 22nd ACM International Conference on Multimedia*, Orlando, Florida, USA, 11 2014.

[3] C. Bhatt, N. Pappas, M. Habibi, and A. Popescu-Belis. Idiap at MediaEval 2013: Search and Hyperlinking Task. In *MediaEval 2013 Workshop*, 2013.

[4] S. Chen, G. J. F. Jones, and N. E. O'Connor. DCU Linking Runs at MediaEval 2013: Search and Hyperlinking Task. In *MediaEval*, 2013.

[5] T. De Nies, W. De Neve, E. Mannens, and R. Van de Walle. Ghent University-iMinds at MediaEval 2013: An Unsupervised Named Entity-based Similarity Measure for Search and Hyperlinking. In *MediaEval 2013 Workshop*, pages 1–2, Oct. 2013.

[6] M. Eskevich, R. Aly, R. Ordelman, S. Chen, and G. J. F. Jones. The search and hyperlinking task at mediaeval 2013. In *MediaEval*, Barcelona, Spain, October 2013.

[7] M. Eskevich, R. Aly, D. N. Racca, R. Ordelman, S. Chen, and G. J. Jones. The Search and Hyperlinking Task at MediaEval 2014. In *Proceedings of MediaEval 2014 Workshop*, 2014.

[8] M. Eskevich, G. J. 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, New York, NY, USA, 2013. ACM.

[9] M. Eskevich and G. J. F. Jones. Time-based Segmentation and Use of Jump-in Points in DCU Search Runs at the Search and Hyperlinking Task at MediaEval 2013. In *MediaEval 2013 Workshop*, 2013.

[10] P. Galuscáková and P. Pecina. at MediaEval 2013 Search and Hyperlinking Task. In *MediaEval 2013 Workshop*, 2013.

[11] P. Galuščáková and P. Pecina. CUNI at MediaEval 2014 Search and Hyperlinking Task: Search task experiments. In *Working Notes Proceedings of the MediaEval 2014 Workshop*, volume 1623 of *Workshop Proceeding*, Barcelona, Spain, 2014. CEUR Workshop Proceeding.

[12] C. Guinaudeau, A.-R. Simon, G. Gravier, and P. Sébillot. HITS and IRISA at MediaEval 2013: Search and Hyperlinking Task. In *MediaEval 2013 Workshop*, 2013.

[13] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of 4th Alvey Vision Conference*, pages 147–151, 1988.

[14] L. Lamel and J.-L. Gauvain. Speech processing for audio indexing. In *Advances in Natural Language Processing (LNCS 5221)*, pages 4–15. Springer, 2008.

[15] P. Lanchantin, P. Bell, M. J. F. Gales, T. Hain, X. Liu, Y. Long, J. Quinnell, S. Renals, O. Saz, M. S. Seigel, P. Swietojanski, and P. C. Woodland. Automatic transcription of multi-genre media archives. In *Proceedings of the First Workshop on Speech, Language and Audio in Multimedia, Marseille, France, August 22-23, 2013*, pages 26–31, 2013.

[16] H. Le, Q. Bui, B. Huet, B. Cervenková, J. Bouchner, E. E. Apostolidis, F. Markatopoulou, A. Pournaras, V. Mezaris, D. Stein, S. Eickeler, and M. Stadtschnitzer. LinkedTV at MediaEval 2014 Search and Hyperlinking Task. In *Proceedings of MediaEval 2014 Workshop*, 2014.

[17] M. Lokaj, H. Stiegler, and W. Bailer. TOSCA-MP at Search and Hyperlinking of Television Content Task. In *MediaEval 2013 Workshop*, 2013.

[18] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[19] R. J. F. Ordelman, M. Eskevich, R. Aly, B. Huet, and G. J. F. Jones. Defining and evaluating video hyperlinking for navigating multimedia archives. In *Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18-22, 2015 - Companion Volume*, pages 727–732, 2015.

[20] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A. F. Smeaton, and G. QuÃĹenot. TRECVID 2012 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proceedings of TRECVID 2012*. NIST, USA, 2012.

[21] A. Rousseau, F. Bougares, P. Deléglise, H. Schwenk, and Y. Estève. LIUM's systems for the IWSLT 2011 Speech Translation Tasks. In *Proceedings of IWSLT 2011*, San Francisco, USA, 2011.

[22] M. Sahuguet, B. Huet, B. Cervenkova, E. Apostolidis, V. Mezaris, D. Stein, S. Eickeler, J. L. Redondo Garcia, and L. Pikora. LinkedTV at MediaEval 2013 search and hyperlinking task. In *MediaEval 2013 Workshop*, Barcelona, Spain, 10 2013.

[23] K. Schouten, R. Aly, and R. Ordelman. Searching and Hyperlinking using Word Importance Segment Boundaries in MediaEval 2013. In *MediaEval 2013 Workshop*, 2013.

[24] P. Sidiropoulos, V. Mezaris, and I. Kompatsiaris. Enhancing Video concept detection with the use of tomographs. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, Melbourne, Australia, September 2013.

[25] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso. Temporal Video Segmentation to Scenes Using High-Level Audiovisual Features. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(8):1163–1177, August 2011.

[26] A. Simon, G. Gravier, P. Sébillot, and M. Moens. IRISA and KUL at mediaeval 2014: Search and hyperlinking task. In *Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Catalunya, Spain, October 16-17, 2014.*, 2014.

[27] K. van de Sande, T. Gevers, and C. Snoek. Evaluating Color Descriptors for Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, September 2010.

[28] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek. Visual Word Ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1271–1283, july 2010.

[29] C. Ventura, M. Tella, and X. Giró-i Nieto. UPC at MediaEval 2013 Hyperlinking Task. In *MediaEval 2013 Workshop*, Barcelona, Catalonia, 10/2013 2013. CEUR Workshop Proceedings Vol-1043, CEUR Workshop Proceedings Vol-1043.

[30] C. Wartena. Comparing segmentation strategies for efficient video passage retrieval. In *10th International Workshop on Content-Based Multimedia Indexing, CBMI 2012, Annecy, France, June 27-29, 2012*, pages 1–6, 2012.

[31] M. Yeung, B.-L. Yeo, and B. Liu. Segmentation of video by clustering and graph analysis. *Computer Vision and Image Understanding*, 71(1):94–109, July 1998.