Automatic Fine-grained Hyperlinking of Videos within a Closed Collection using Scene Segmentation

Evlampios Apostolidis
CERTH-ITI
Thessaloniki, Greece
apostolid@iti.gr

Vasileios Mezaris
CERTH-ITI
Thessaloniki, Greece
bmezaris@iti.gr

Mathilde Sahuguet
EURECOM
Sophia Antipolis, France
sahuguet@eurecom.fr

Benoit Huet
EURECOM
Sophia Antipolis, France
huet@eurecom.fr

Barbora Červenková
University of Economics
Prague, Czech Republic
cervenkovab@gmail.com

Daniel Stein
Fraunhofer IAIS
Sankt Augustin, Germany
daniel.stein@iais.fraunhofer.de

ABSTRACT

This paper introduces a framework for establishing links between related media fragments within a collection of videos. A set of analysis techniques is applied for extracting information from different types of data. Visual-based shot and scene segmentation is performed for defining media fragments at different granularity levels, while visual cues are detected from keyframes of the video via concept detection and optical character recognition (OCR). Keyword extraction is applied on textual data such as the output of OCR, subtitles and metadata. This set of results is used for the automatic identification and linking of related media fragments. The proposed framework exhibited competitive performance in the Video Hyperlinking sub-task of MediaEval 2013, indicating that video scene segmentation can provide more meaningful segments, compared to other decomposition methods, for hyperlinking purposes.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Retrieval models; H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing

Keywords: Video hyperlinking; scene segmentation; concept detection; keyword extraction; data indexing

1. INTRODUCTION

Nowadays there is a constantly increasing volume of media on the web, available either on video-sharing web-sites such as YouTube and Vimeo, or within multimedia collections of content providers such as web-based broadcasters and news organizations. A very natural paradigm of exploring this content includes a user that watches a video and then needs to follow up on an interesting fragment of it (denoted "anchor" or "context" in sequel), by easily navigating to other videos or video fragments on the same topic, featuring the same persons, places or objects. The goal of video hyperlinking is to provide links to the user for accessing related content, thus supporting and facilitating the navigation within a multimedia collection. This is by no means a trivial task, particularly at a fine-granularity level (fragment-level). Text-based search and retrieval strategies can be used only when textual information has already been assigned to the videos by human annotators, and even so, such annotations most often briefly describe each entire video rather than specifically the relevant segments of it. This highlights two main weaknesses of the current approaches for establishing such links between multimedia content items: the definition of semantically meaningful and coherent parts of the content using simple temporal segmentation strategies that rely either on low-level visual or textual data is not sufficient, while due to the known semantic gap, automatic hyperlinking based on text features only fails to exploit the rich multi-modal content of videos.

Aiming to address these problems, we propose a framework that creates story-telling multimedia fragments from videos, and exploits the high-level semantics of the visual stream for assessing the relevance between media fragments and for establishing links whenever is appropriate. Using the created links a viewer can efficiently navigate, when seeking relevant content within a multimedia collection, according to the vision of video hyperlinking.

MM'14, November 3–7, 2014, Orlando, Florida, USA.

http://dx.doi.org/10.1145/2647868.2655041.

2. RELATED WORK

Several approaches for video hyperlinking have been proposed and evaluated in the last couple of years, mainly in the scope of the Hyperlinking sub-task of the MediaEval benchmarking activity. The relevant literature includes methods that either rely only on textual or visual information, or combine these different modalities using various fusion and re-ranking schemes, for linking related media fragments.

A unimodal text-based method that combines fixed-length segmented transcripts (created by automatic speech recognition (ASR)), enriched with extracted entities, and an unsupervised similarity metric for identifying relevant media fragments, was introduced in [8]. A similar approach that evaluates similarity on the video-level and a more detailed segment-level using partitioned ASR transcripts and extracted entities was presented in [5], while another method that

defines fragments by performing a term-based speech segmentation was introduced in [11]. The algorithm of [7] creates a short-list of semantically related videos based on ASR transcripts, and then defines the appropriate fragments by constructing word similarity graphs, or by evaluating the lexical cohesion. Another unimodal approach based exclusively on visual information was proposed in [16], where the visual similarity between shots is assessed by representing the keyframes of the corresponding shots using a Bag-of-Words model built from local descriptors.

Alternatively, multimodal hyperlinking methods that combine text- and visual-based analysis were also proposed. In [10], similar media fragments are defined by building a graph based on keywords extracted from ASR transcripts, while visual similarity is assessed with the help of local descriptors. In [4], the video is initially segmented into topics based on ASR transcripts and subtitles, and a list with the most relevant segments to a given "anchor" is defined via cosine similarity. Visual similarity between these segments and the "anchor" is then assessed using visual concepts, and the re-ranked list of segments is the final output of the algorithm. A similar approach was introduced in [3], where the textual similarity is computed based on matching terms and named entities, and the visual similarity is estimated by extracting local descriptors from each 5-th frame of the video segments.

Textual information extracted from subtitles, ASR transcripts or metadata, can be a good indicator about the relevance between entire videos. However, the definition of appropriate media fragments within videos that are closely related to the human information needs is a key goal of video hyperlinking. Text-only approaches for specifying these segments lack the information included in the visual stream, often over-segmenting the media items, while visual shot segments are too fine-grained for describing an entire story telling part of the video. Aiming to address this problem, differently from any other work on this topic, we propose a video hyperlinking framework where the elementary temporal segments are automatically formed by a method for the segmentation of video in scenes, which are subsequently annotated by using a rich set of audio-visual analysis techniques.

3. PROPOSED APPROACH

This section describes the overall architecture of the proposed system, focusing further on the scene segmentation analysis module which defines media fragments that are used as starting and ending points of the hyperlinking procedure.

3.1 Overall hyperlinking framework

As shown in Fig. 1, the proposed framework consists of two main components. The analysis component includes all the utilized analysis modules, and can be further divided in a group of visual-based techniques and a method that processes textual information. The storage component contains the data storage and indexing structures that facilitate the retrieval and linking of related media fragments.

The off-line processing of the multimedia collection starts by decomposing the videos into shots (e.g., using [2]) and extracting, for each shot, a number of representative keyframes that will be used as input in subsequent visual analysis techniques. Then, the scene segmentation technique of [13] groups the detected shot segments into bigger story-telling units (scenes), aiming to define a more meaningful frag-

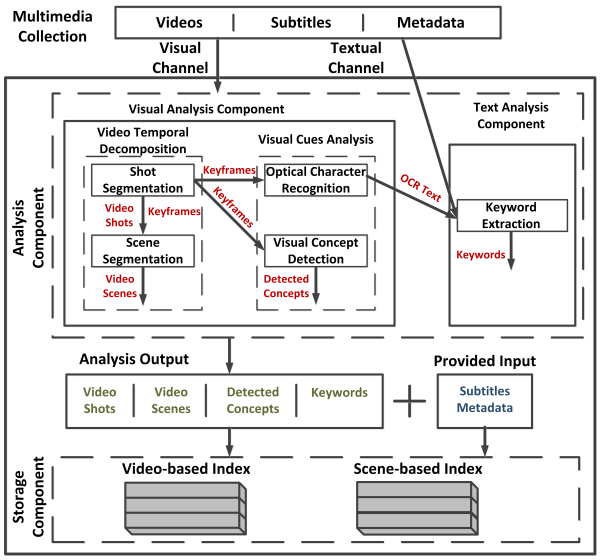

Figure 1: The proposed framework for multi-modal analysis and indexing, that supports the retrieval and hyperlinking of media fragments.

mentation of the video. Using the keyframes, visual cues are extracted by performing OCR and visual concept detection. OCR is based on the tesseract engine[1] and the text localization algorithm presented in [14]. Visual concepts are detected from each keyframe by applying the concept detection algorithm of [12], which uses a sub-set of 10 base detectors per concept and a set of 151 semantic concepts selected from the list of concepts defined in the TRECVID 2012 SIN task [9]. OCR results and any other available textual resources (e.g., the videos' subtitles and metadata) are fed to the utilized keyword extraction algorithm, which is the one presented in [15]. The output of this analysis is a number of keywords and visual concepts which are assigned to different video fragments.

The produced analysis results, along with the subtitles and metadata of the videos, are indexed using the storage component of the framework. This is based on the Solr/Lucene platform and creates indexes that contain data at two granularities: the video level and the scene level.

3.2 Temporal video segmentation into scenes

As already mentioned, starting from a decomposition of each video of the multimedia collection into shots, the proposed framework defines a more meaningful video segmentation into story-telling parts. For this, the scene segmentation algorithm of [13] is utilized. This method groups shots into sets that correspond to individual scenes of the video, based on content similarity and temporal consistency among shots. Content similarity in our experiments means 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 well known Scene Transition Graph (STG) algorithm [17]. The first extension reduces the computational cost of STG-based shot grouping by consid-

[1] `http://code.google.com/p/tesseract-ocr/`

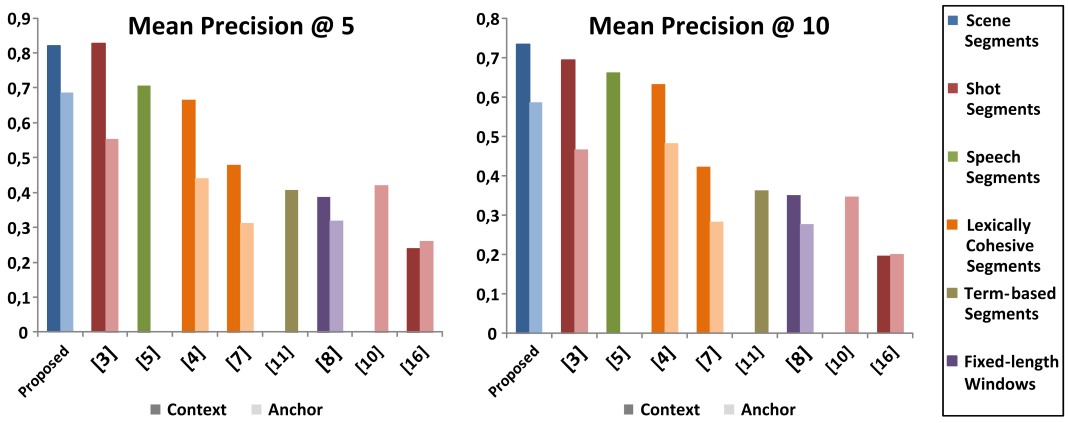

Figure 2: The best mean performance of each participating approach in the Hyperlinking sub-task of MediaEval 2013, also in relation to the segmentation unit employed by each team (see legend on the right).

ering shot linking transitivity and the fact that scenes are by definition convex sets of shots. The second one builds on the former to construct a probabilistic framework that alleviates the need for manual STG parameter selection. Based on these extensions, and as reported in [13], the applied technique is capable of identifying automatically the scene-level structure of videos belonging to different genres (e.g., documentaries, movies), providing results that match well the human expectations, while the needed processing time is only a very small fraction ($< 1\%$) of the overall analysis.

4. EXPERIMENTAL RESULTS

The performance of the proposed framework was evaluated on the Hyperlinking sub-task of the MediaEval 2013 benchmarking activity [6]. The used dataset is made of 1667 hours of video (2323 videos from BBC) of various content, such as news shows, talk shows, sitcoms and documentaries. For each video the organizers provided also manually transcribed subtitles, ASR transcripts, textual metadata and automatically detected shot boundaries and keyframes.

The goal of the Hyperlinking sub-task is to search a multimedia collection for content related to a given media fragment. Specifically, the task organizers defined a set of 30 "anchors" (media fragments described by the video's name and their start and end times; thus, no further temporal segmentation of them is necessary), which are used as the basis for seeking related content within the provided collection. For each "anchor", a broader yet related temporal segment with contextual information about the "anchor", called "context", was also defined. For evaluating the hyperlinking performance, Precision @ k (which counts the number of relevant fragments within the top k of the ranked list of hyperlinks, with k being 5 and 10) was used. Moreover, three slightly different functions were defined for measuring the relevance of a retrieved segment; the "overlap relevance", which considers the temporal overlap between a retrieved segment and the actual one; the "binned relevance", which assigns segments into bins; and the "tolerance to irrelevance", which takes into account only the start times of the retrieved segments [1].

Given a pair of "anchor" and "context" fragments the proposed framework initially creates automatically two queries, one using only the "anchor" information and another one using both "anchor" and "context" information, which are going to be applied on the created indexes (Section 3.1). These queries are defined by extracting keywords from the subtitles of the "anchor"/"context" fragments, and by applying visual concept detection. The latter is performed on the keyframes of all shots of the corresponding media fragment and its results are combined using max pooling (i.e., keeping for each concept the highest confidence score). Our framework then applies these queries on the video-level index; this step filters the entire collection of videos, resulting in a much smaller set of potentially relevant videos. Using the new limited set of videos, the same queries are applied on the scene-level index, and a ranked list with the scenes that were identified as the most relevant ones is returned, forming the output of the proposed system.

Figure 2 illustrates the best mean performance of each participating team in MediaEval 2013 (in terms of Precision @ k using the "overlap relevance" metric) when only the "anchor" or the "anchor" and "context" information is exploited, also indicating (by color) which were the segmentation units utilized by each approach. As shown, when only the "anchor" is known, our proposed approach exhibits the highest performance for k equal to 5 or 10, while it is among the top-2 highest performers when "context" information is also included. Specifically, the k-th first items (hyperlinks) proposed by our system to the user are very likely to include the needed media fragment (over 80% for the top-5 and over 70% for the top-10). Moreover, the comparison of the different video decomposition approaches shows that the visual-based segmentation techniques (scene or shot segmentation) are more effective than other speech-based, text-based or fixed-window segmentation methods.

The competitiveness of the developed hyperlinking approach is also highlighted in Table 1. This table contains the best scores of each participating team for the Mean Precision @ 5 and @ 10 measures, according to the different defined relevance functions (highest scores are in bold font; a dash means that no run was submitted to MediaEval 2013). As shown, the proposed framework achieves the best performance in 11 out of 12 cases.

We also ran an experiment with a variation of our approach that used a simple temporal window (defined by

Table 1: The best Mean Precision @ 5 and @ 10 scores (for the different relevance measures) for the teams participating to the Hyperlinking sub-task of MediaEval 2013, using "anchor" and "context" information.

	Mean Precision @ 5						Mean Precision @ 10					
	Overlap Relevance		Binned Relevance		Toll. to Irrelevance		Overlap Relevance		Binned Relevance		Toll. to Irrelevance	
	Context	Anchor	Context	Anchor	Context	Anchor	Context	Anchor	Context	Anchor	Context	Anchor
Proposed	0,8200	**0,6867**	**0,7200**	**0,6600**	**0,6933**	**0,6133**	**0,7333**	**0,5867**	**0,6333**	**0,5467**	**0,6367**	**0,5133**
[3]	**0,8267**	0,5533	0,5400	0,5333	0,5933	0,5133	0,6933	0,4667	0,3867	0,4333	0,4433	0,4200
[5]	0,7067	-	0,6000	-	0,5333	-	0,6633	-	0,5667	-	0,4667	-
[4]	0,6667	0,4400	0,6333	0,5000	0,4600	0,3867	0,6333	0,4833	0,5167	0,4867	0,4433	0,4033
[7]	0,4800	0,3133	0,4600	0,3400	0,4667	0,3133	0,4233	0,2833	0,4100	0,3000	0,4100	0,2733
[11]	-	0,4067	-	0,3933	-	0,3600	-	0,3633	-	0,3500	-	0,3267
[8]	0,3867	0,3200	0,3867	0,3267	0,3667	0,3067	0,3500	0,2767	0,3500	0,2800	0,3233	0,2600
[10]	-	0,4200	-	0,4000	-	0,3400	-	0,3467	-	0,3267	-	0,2900
[16]	0,2400	0,2600	0,2400	0,2600	0,2333	0,2467	0,1967	0,2000	0,1967	0,1900	0,1933	0,1900
Windowing	0,5733	0,4467	0,6067	0,5000	0,4600	0,3467	0,4833	0,3200	0,5333	0,3733	0,4000	0,2533

grouping shots that are no more than 10 sec. apart) for determining the temporal segments used for hyperlinking, instead of the outcome of scene segmentation (last row of Table 1). The comparison again indicates that automatically detected scenes are more meaningful video fragments for hyperlinking, compared to simpler temporal segmentations (e.g., windowing).

5. CONCLUSIONS

In this paper we presented an approach for analysing multimedia collections aiming to define meaningful media fragments and create links between related content. The proposed framework consists of a set of multi-modal analysis techniques, which include methods for video segmentation and for the identification of visual cues, as well as an algorithm for extracting keywords from textual data. Our participation in the Hyperlinking sub-task of MediaEval 2013 and related experiments showed that the proposed framework performs very well, and highlighted the importance of video scenes in video hyperlinking tasks.

6. ACKNOWLEDGMENTS

This work was supported by the European Commission under contract FP7-287911 LinkedTV.

7. ADDITIONAL AUTHORS

Additional authors: Stefan Eickeler (Fraunhofer IAIS, email: `stefan.eickeler@iais.fraunhofer.de`), José Luis Redondo García (Eurecom, email: `redondo@eurecom.fr`), Raphaël Troncy (Eurecom, email: `troncy@eurecom.fr`) and Lukás Pikora (University of Economics Prague, email: `lukas.pikora@vse.cz`).

8. REFERENCES

[1] R. Aly, M. Eskevich, R. Ordelman, et al. Adapting binary information retrieval evaluation metrics for segment-based retrieval tasks. Technical report, ArXiv e-prints, 2013.

[2] E. Apostolidis and V. Mezaris. Fast shot segmentation combining global and local visual descriptors. In *Proc. IEEE ICASSP*, Florence, Italy, May 2014.

[3] W. Bailer, M. Lokaj, and H. Stiegler. Context in video search: Is close-by good enough when using linking? In *ACM ICMR*, Glasgow, UK, April 1-4 2014.

[4] C. A. Bhatt, N. Pappas, M. Habibi, et al. Multimodal reranking of content-based recommendations for hyperlinking video snippets. In *ACM ICMR*, Glasgow, UK, April 1-4 2014.

[5] S. Chen, G. J. F. Jones, and N. E. O'Connor. DCU linking runs at MediaEval 2013: Search and Hyperlinking task. In *MediaEval*, 2013.

[6] M. Eskevich, R. Aly, R. Ordelman, et al. The Search and Hyperlinking task at MediaEval 2013. In *MediaEval*, 2013.

[7] C. Guinaudeau, A.-R. Simon, G. Gravier, et al. HITS and IRISA at MediaEval 2013: Search and Hyperlinking task. In *MediaEval*, 2013.

[8] T. D. Nies, W. D. Neve, E. Mannens, et al. Ghent University-iMinds at MediaEval 2013: An unsupervised named entity-based similarity measure for search and hyperlinking. In *MediaEval*, 2013.

[9] P. Over, G. Awad, M. Michel, et al. TRECVID 2012 – An overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proc. of TRECVID 2012*. NIST, USA, 2012.

[10] J. Preston, J. S. Hare, S. Samangooei, et al. A unified, modular and multimodal approach to search and hyperlinking video. In *MediaEval*, 2013.

[11] K. Schouten, R. Aly, and R. Ordelman. Searching and Hyperlinking using word importance segment boundaries in MediaEval 2013. In *MediaEval*, 2013.

[12] P. Sidiropoulos, V. Mezaris, and I. Kompatsiaris. Enhancing video concept detection with the use of tomographs. In *Proc. IEEE ICIP*, 2013.

[13] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, et al. Temporal video segmentation to scenes using high-level audiovisual features. *IEEE Trans. on CSVT*, 21(8):1163–1177, Aug. 2011.

[14] D. Stein, S. Eickeler, R. Bardeli, et al. Think before you link – Meeting content constraints when linking television to the web. In *NEM Summit 2013, 28-30 October 2013, Nantes, France*.

[15] S. Tschöpel and D. Schneider. A lightweight keyword and tag-cloud retrieval algorithm for automatic speech recognition transcripts. In *Proc. 11th Annual Conf. of the Int. Speech Communication Association*, 2010.

[16] C. Ventura, M. Tella-Amo, and X. G. i Nieto. UPC at MediaEval 2013 Hyperlinking task. In *MediaEval*, 2013.

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