Convenient Discovery of Archived Video Using Audiovisual Hyperlinking

Roeland J.F. Ordelman
University of Twente &
Netherlands Institute for
Sound and Vision
The Netherlands

Robin Aly
Data Management
University of Twente
The Netherlands

Maria Eskevich
EURECOM
Sophia Antipolis
France

Benoit Huet
EURECOM
Sophia Antipolis
France

Gareth J.F. Jones
ADAPT Centre / CNGL
School of Computing
Dublin City University, Ireland

ABSTRACT

This paper overviews ongoing work that aims to support end-users in conveniently exploring and exploiting large audiovisual archives by deploying multiple multimodal linking approaches. We present ongoing work on multimodal video hyperlinking, from a perspective of unconstrained link anchor identification and based on the identification of named entities, and recent attempts to implement and validate the concept of outside-in linking that relates current events to archive content. Although these concepts are not new, current work is revealing novel insights, more mature technology, development of benchmark evaluations and emergence of dedicated workshops which are opening many interesting research questions on various levels that require closer collaboration between research communities.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Query formulation, Search process*

General Terms

Human factors; Experimentation; Performance

Keywords

Audiovisual Access; Video Hyperlinking; Multimodal Search

1. INTRODUCTION

Online archives of multimodal content, including broadcast and cultural heritage are growing very rapidly. These archives have considerable social and economical potential, and are often the result of considerable national investment in digitization and storage of the content they hold. As such, there is a need for new technologies to stimulate active appropriation [15] – the use of existing digital audiovisual resources by users or user communities according to their expectations, needs, interests or desires. The traditional access model assumes that end-users visit a locally maintained search portal and use keyword search and advanced filter options to fulfill their 'information need'. Such a model is too limited to achieve effective use, social impact and explore the full range of potentially relevant content within big video archives. Indeed it is not that clear whether public end-users are generally aware of the existence of such portals, or when they are, if they are aware of their potential, and if so, if access to them is sufficiently *convenient*: readily accessible online, do they contain the necessary information and are they easy to use, and can the information be accessed quickly [6].

Results of an evaluation of a video search system, that we recently carried out to investigate the performance of state-of-the-art visual search [16], indicate that end-users do not have much clue of the potentially interesting footage available in well-established broadcast archives. Confronted with a search system for large video collections of BBC content and material from the Netherlands Institute for Sound and Vision, end-users typically started with a few searches about a celebrity, their own home town, or a hobby or personal interest, and then often quickly stalled with the question to the experimenters: 'what else might there be in the archive'?

With the clear limitations of existing tools and the demands and opportunities to realize the potential value of this material, we have recently been experimenting with a number of alternative models for accessing content in big audio-visual archives. These have focused on increasing user awareness of their potential, and convenience of content access. An example of such a model is the 'random access' model resembling Google's 'Feeling Lucky' option based on thousands of (pre-trained) visual concepts automatically detected in the archive. By shaking the tablet or phone on which the application runs, a new set of visual concepts (categories, locations, persons, events) are shown to the user in a list format that can be easily scrolled through vertically

SLAM'15 October 30, 2015, Brisbane, Australia
ACM 978-1-4503-3749-6/15/10.
http://dx.doi.org/10.1145/2802558.2814652.

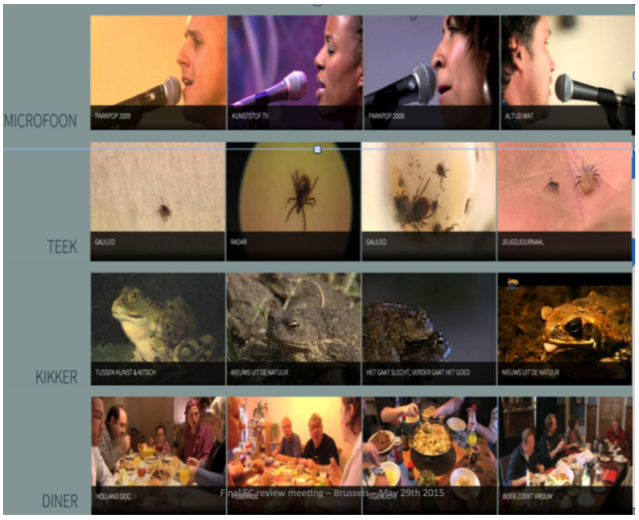

Figure 1: 'Random' archive access using visual concept detectors. Scrolling down gives more visual concepts, scrolling to the right gives more media fragments for the visual concept.

(concept scrolling) and horizontally (fragment scrolling) as depicted in Figure 1. This model provides end-users with insights into archival footage and stimulates serendipitous browsing through its content, perfectly suiting an entertainment type of use of a big archive.

To stimulate alternative use scenarios, we have been looking particularly into access models to stimulate discovery in a convenient manner based on the concept of linking: redirecting users from one multimedia information source to the other. Looking at linking from an archive point of view, we have distinguished three types: (i) inside-in video-to-video linking (further referred to as 'video hyperlinking') that recommends topically related media fragments based on properties of a fragment the user is currently watching, (ii) outside-in multimedia linking that takes a multimedia information source in the outside world and recommends archival media fragments for that, and (iii) inside-out multimedia linking that recommends multimedia information sources in the outside world based on properties of a media fragment the user is currently watching or listening to. In this paper we overview our recent work on these access models, as they are expected to be the most suitable to stimulate the exploration of audiovisual archives.

2. VIDEO HYPERLINKING

In automatic video-to-video hyperlinking, we envisage an archive as consisting of an infinite set of media fragments that can be linked to other media fragments in the archive based on the available multimodal information in the fragments. A video hyperlinking system capable of automatically creating links resembles a video search system: it extracts a query representation from an anchor media fragment, applies this to a video search system, and identifies potentially relevant fragments, targets, to present to the user.

We envisage a typical use case being navigation through large quantities of archived video content via a multi dimensional link structure at the level of media fragments [3, 7]. This navigation can have an exploratory nature where links could be ranked following a PageRank like structure, or be part of a storytelling approach that assembles related media fragments on the basis of a topic. Thus, the goal of searching in a video hyperlinking context is to find content that is *about* what is represented in the anchor – we sometimes refer to this as 'topically related' – and not content that is *based upon* it, which is *similar* to it, or has identical semantic labels. One important implication of this is that the context of the anchor in the 'about' case is of significant importance.

The automatic video hyperlinking process is composed of the following steps: (i) *anchor identification*, in which for a given video a set of anchor-segments are determined, identified by their start time and end time, for which users might benefit from a link to related information, (ii) *anchor representation*, in which a query input for the video search system is created from the segment (and potentially other information), (iii) *target search*, in which a ranked list of video segments, defined by their start time and end time, is produced by the video search system in response to the query, and (iv) *target presentation*, in which elements from the ranked list of search results are presented to the user.

In our recent work we have been investigating an unconstrained, multimodal perspective on the identification of anchor points, and a perspective based on the detection of entities based on existing metadata and/or automatic audiovisual analysis. The first in the context of two benchmark evaluations related to video hyperlinking (TRECVid and MediaEval), and the second in the context of a European research project on Linked Television [14].

2.1 Unconstrained anchor identification

The foregoing discussion assumes that the location of source anchors for the video hyperlinking process are known. However, how to identify such anchors is itself an open and challenging question. Ideally we want such anchors to correspond to those of interest to users of the content. However, the concept of video hyperlinking is not something with which current users are generally familiar. This impacts studies of user behaviour in video hyperlinking scenarios: selecting 'interesting' segments in a video to create anchors is completely new to current users. Also, they are generally unfamiliar with the concept of multi-modality as applied in video search. In the user experiments we conducted in the course of developing test collections for video hyperlinking benchmark evaluations, we therefore put a lot of effort into explaining the concepts related to video hyperlinking. Although participants claimed to understand these concepts, it was our experience that they needed some time to become familiar with the idea of identifying suitable anchors.

Previous work has also suggested that it may be difficult for users to identify hyperlinks in material that they are not genuinely interested in [1]. Thus instead of providing our users with a video from which to select anchors, we placed it within the context of a search scenario in which anchor identification took place as a second step (see [2] for a more elaborate description). For the search part we provided them with a prototype audio-visual search system [16] that supports both visual and textual search. Also we gave our users a monetary compensation for their efforts to improve their intrinsic involvement.

For a given relevant clip found in the search process, we asked the user to identify anchors within it from which they would want to link to related video content. To do this they were provided with a virtual cutter tool which enabled them to define a start and an end point for a video fragment. We also asked the user to give a textual description of what was

contained in the anchor (e.g. "about a world famous singer and his relation with organized crime"), main characteristics of the anchor as starting point for linking –the whole scene, the speech, a moving object, a static object, or the category 'other' that could be used for music as the main characteristic for an anchor–, and a description of what they expected to see in the link targets (e.g. "mafia clips; connections between mafia and other singers/famous people").

Interestingly we found that participants created anchors that referred primarily to spoken content and whole scenes, anchors referring to visual objects were under-represented. This user focus on spoken content is slightly at odds with the common research focus on the visual dimension of audiovisual archives, and suggests that there should be greater research focus on the use of the audio stream and true multimodality in query processing and search. Also, we learned from interviews with the participants that there can be differences in the underlying intention of manual anchor creation either from the perspective of *a content producer* generating anchors s/he thinks an end-user would be interested in, or from the perspective of *an end-user* wanting a certain anchor in the context of watching a video clip. In future user studies we will therefore focus more on the mode that users are in while defining potential anchors.

Of course, in operation, we would hope to automate the process of anchor identification, as well the query construction and link identification processes. In order to do this, we need to study the characteristics of manually selected links and seek to develop algorithmic methods to replicate expected human selections based on the available multimodal features in the content archive.

When an anchor is defined the multimodal information within the start and end time of the fragment needs to be translated into a query suitable for use in a video search system. The query may be based on available textual features (metadata, speech transcripts) and visual features (keyframes, temporal events) in the anchor video. An anchor may also incorporate context information such as its source video or local information within this video Given that we assume video hyperlinking to operate on large video collections, the information contained within the video will typically be extracted automatically using audio (e.g., speech recognition, speaker recognition) and video analysis tools that will introduce noise in the extracted features.

For TRECVid 2015 Video Hyperlinking task[1] 100 anchors were created by the participants in our study to be used as test anchors where the goal is to return for each anchor a ranked list of hyperlinking targets in a dataset of some 3000 hours of BBC video content. The relevance of the targets will be evaluated using crowd sourcing via the Amazon MTurk platform [2]. Results of the benchmark will be presented at the TrecVID workshop in the fall of 2015. A set of video fragments with anchor labels will be used as ground truth for the automatic anchoring task in the MediaEval 2015 benchmark on Search and Anchoring in Video Archives.

2.2 Entity-based anchor identification

The entity-based anchor identification perspective resembles a wikification scenario: entity linking with Wikipedia as the target knowledge base [10]. This scenario fits a use case of content producers generating video hyperlinks. The type of technique is targeted at the richer interactive television experiences in which broadcast companies are becoming increasingly interested. In this setting television content is enriched with hyperlinks that connect the primary content to other data sources that are intended to enhance the attractiveness of watching television in either a linear fashion or on-demand. By default, such links are currently created manually by broadcast companies' editorial departments. However, the identification of media-fragments that can be used as anchors in the primary content, and the selection of appropriate link targets completely manually is labour intensive. Automating at least parts of the hyperlink generation process is essential to be able to provide users with a rich set of links. A second argument for incorporating technology into the link generation process is that manual hyperlink generation is inherently limited by a human editor's subjective view on anchor selection and limited knowledge of relevant target videos in big audiovisual archives. Not every anchor candidate may be useful or lead to interesting content in automated video hyperlinking. To enable semi-supervised hyperlink generation, we created video hyperlink editors, such as VideoHypE [4], and LinkedTV Editor Tool (ET) [3]. Manual intervention can take place on several levels: the editor can accept and reject hyperlinks or add hyperlinks that are not provided by the system.

3. OUTSIDE-IN LINKING

Connecting current events published in online textual media with the archival footage is another linking approach to stimulate discovery and re-use of archival footage. Instead of the 'random seeds' of visual concept detectors described in the introduction section, event descriptions such as an RSS news feed or a blog post, can be used to obtain pointers to interesting media fragments to guide users into the archive and foster its exploration and exploitation. This approach is interesting for both public users and professionals – for the former as a means to reduce effort by providing a form of 'query-free search' [8], and for the latter in a scenario that requires fast response and broad contextual coverage of incoming news events.

The outside-in linking concept is not new. Linking related events in different media types was studied among others in [5, 11, 13] for example to enrich sparsely annotated data in the one media with textually rich content from the other. The novelty in our work is that thanks to improving techniques and computation facilities, new opportunities arise for convenient discovery of rich archival content (see also [12] for a more extensive overview). Most notable with respect to improving techniques for linking in recent years, is the boost in quality of visual concept detection (faces, visual categories, locations, events) that justifies a true multimodal approach. Such an approach was recently implemented in a research prototype based on RSS news feeds where extracted named entities from the feeds were matched with archive metadata, speech transcripts and the outputs from visual analysis on the video collection (Figure 2).

In a qualitative evaluation [9], public end-users expressed their interest in the concept, but also indicated that the quality of the recommendations was unsatisfactory now and then. They also noted that the concept would in itself not

[1] http://www-nlpir.nist.gov/projects/tv2015/tv2015.html#lnk

[2] www.mturk.com

[3] http://editortool.linkedtv.eu/

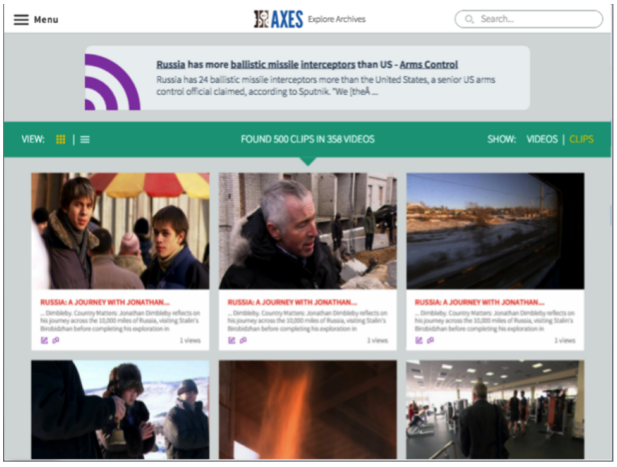

Figure 2: Online textual media such as news feeds are used to obtain pointers to related archival footage via a multimodal linking approach.

attract them to a search portal, but that it would be an interesting feature for current events sites that they are already using frequently. An informal analysis pointed to two important factors for quality issues: the fine-tuning of the structured query formulation process from news feeds texts and data sparseness given the relatively small data set (3000 hours) that was used.

4. SUMMARY AND CONCLUSION

This paper presented a brief overview of ongoing work aimed towards improving the capabilities for exploring, and with that, unlocking the social and economical potential of large audiovisual archives, and deploying the concept of multimodal linking. Although the concept of linking multimedia content is not new, we feel that over the last years, the community has brought this topic to a next level, both research-wise bringing new insights and more mature technology, and community-wise, among others by establishing benchmark evaluations and dedicated workshops. This next level opens up many interesting research questions and also opportunities, which need collaboration between multiple research communities, to bring together adjacent research communities, for example with respect to individual facets of multimodality and their combination or linked open data.

5. ACKNOWLEDGMENTS

This work was supported by the European Commission's 7th Framework Programme (FP7) under FP7-ICT 269980 (AXES) and FP7-ICT 287911 (LinkedTV), the Dutch national program COMMIT/; Science Foundation Ireland (Grant No 12/CE/I2267) as part of the Centre for Next Generation Localisation (CNGL) project at DCU; Bpifrance within the NexGen-TV Project, under grant number F1504054U. The user studies were executed in collaboration with Jana Eggink and Andy O'Dwyer from BBC Research.

References

[1] R. Aly, K. McGuinness, M. Kleppe, R. Ordelman, N. E. O'Connor, and F. de Jong. Link anchors in images: Is there truth? In *Proceedings of the 12th Dutch Belgian Information Retrieval Workshop (DIR 2012)*, pages 1–4, Ghent, 2012. University Ghent.

[2] R. Aly, R. Ordelman, M. Eskevich, G. J. F. Jones, and S. Chen. Linking inside a video collection - what and how to measure? In *Proceedings of the 22nd International Conference on World Wide Web Companion, IW3C2 2013, Rio de Janeiro, Brazil*, pages 457–460, Brazil, May 2013. ACM.

[3] 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, USA, 11 2014.

[4] J. Blom. Deliverable 1.5, linkedtv annotation tool, final release. Public deliverable, LinkedTV Project (FP7-ICT grant agreement no 287911), 2015.

[5] M. Bron, B. Huurnink, and M. de Rijke. Linking archives using document enrichment and term selection. In S. Gradmann, F. Borri, C. Meghini, and H. Schuldt, editors, *Research and Advanced Technology for Digital Libraries*, volume 6966 of *Lecture Notes in Computer Science*, pages 360–371. Springer Berlin Heidelberg, 2011.

[6] L. S. Connaway, T. J. Dickey, and M. L. Radford. "If it is too inconvenient I'm not going after it": Convenience as a critical factor in information-seeking behaviors. *Library & Information Science Research*, 33(3):179 – 190, 2011.

[7] M. Eskevich, H. Nguyen, M. Sahuguet, and B. Huet. Hyper video browser: Search and hyperlinking in broadcast media. In *ACMMM 2015, 23nd ACM International Conference on Multimedia*.

[8] P. E. Hart and J. Graham. Query-free information retrieval. *IEEE Intelligent Systems*, (5):32–37, 1997.

[9] M. Kleppe and J. Briggeman. Deliverable 1.8, final use case evaluation report. Public deliverable, AXES Project (FP7-ICT grant agreement no 269980), 2015.

[10] 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.

[11] J. Morang, R. J. F. Ordelman, F. M. G. de Jong, and A. J. van Hessen. InfoLink: analysis of Dutch broadcast news and cross-media browsing. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME 2005)*, pages 1582–1585, Los Alamitos, 2005. IEEE Computer Society.

[12] D. W. Oard, A. S. Levi, R. L. Punzalan, and R. Warren. Bridging communities of practice: Emerging technologies for content-centered linking. In *Museums and the Web*, 2014.

[13] D. Stein, E. Apostolidis, V. Mezaris, N. de Abreu Pereira, J. Müller, M. Sahuguet, B. Huet, and I. Lasek. Enrichment of news show videos with multimodal semi-automatic analysis. In *NEM-Summit 2012, Networked and Electronic Media*, Istanbul, Turkey, 10 2012.

[14] D. Stein, A. Öktem, E. Apostolidis, V. Mezaris, J. L. Redondo García, R. Troncy, M. Sahuguet, and B. Huet. From raw data to semantically enriched hyperlinking: Recent advances in the LinkedTV analysis workflow. In *NEM Summit 2013, Networked & Electronic Media*, Nantes, France, 10 2013.

[15] P. Stockinger. *Audiovisual Archives*. John Wiley & Sons, Inc., 2013.

[16] T. Tommasi and R. Aly and K. McGuinness and K. Chatfield and R. Arandjelovic and O. Parkhi and R. Ordelman and A. Zisserman and T. Tuytelaars. Beyond metadata: searching your archive based on its audio-visual content. In *IBC 2014*, Amsterdam, The Netherlands, 2014.