Linking Inside a Video Collection - What and How to Measure?

Robin Aly[1], Roeland J.F. Ordelman[1,2], Maria Eskevich[3], Gareth J.F. Jones[3], Shu Chen[3,4]
[1] Human Media Interaction, University of Twente, Enschede, The Netherlands
[2] Netherlands Institute for Sound and Vision, The Netherlands
[3] Centre for Next Generation Localisation, School of Computing, Dublin City University, Dublin 9, Ireland
[4] CLARITY, School of Electronic Engineering, Dublin City University, Dublin 9, Ireland

ABSTRACT

Although linking video to additional information sources seems to be a sensible approach to satisfy information needs of user, the perspective of users is not yet analyzed on a fundamental level in real-life scenarios. However, a better understanding of the motivation of users to follow links in video, which anchors users prefer to link from within a video, and what type of link targets users are typically interested in, is important to be able to model automatic linking of audiovisual content appropriately. In this paper we report on our methodology towards eliciting user requirements with respect to video linking in the course of a broader study on user requirements in searching and a series of benchmark evaluations on searching and linking.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems—*Human factors*

General Terms

User study, Video linking

1. INTRODUCTION

Automatically linking audiovisual content to other multimedia sources is a topic that has been explored in the context of various use scenarios in the past decade: as an approach towards interactive *non-linear access* to video allowing users to generate narratives on-the-fly by following links in a video (e.g., [13], [14], [10]), as a method to *improve entertainment value* by enriching mono-media audio content with related imagery [8], and especially as a means to *explore additional information sources* while accessing content in a linear fashion [7], [9], or in a search scenario [11],[3],[16]. Although linking seems a sensible approach intuitively, using linked video to satisfy information needs of users is not yet analysed on a fundamental level with users in real-life scenarios.

Essentially, we do not know very well why users would engage into a video link scenario and what users want with respect to the choice of elements in a video that should be linked to additional sources, the so-called *link achors*, and their *link targets*. Although users may be familiar with the concept of hyperlinking in general based on their experience with linking in webpages on the World Wide Web, our exploratory experiments [2] indicate that among others due to the difference in modality, linking in video has its own peculiarities. For example, the selection of a particular video segment for anchoring –that we refer to as anchor segmentation– is based upon audiovisual cues instead of words. This implies that anchoring is more heterogeneous in video then in text and may as well result in ambiguities with respect to its boundaries and contents. A better understanding of anchor segmentation in connection with the motivation of users to follow links in video is crucial to model automatic anchor selection and link generation properly.

Assuming that users have a preconceived information need (in a search scenario) or an emerging one during (linear) access that defines the linked video scenario, the appropriatness of anchors selection and the relevance of link targets could be modelled as a function of this information need. On the other hand, it is well-known that serendipity can be a strong driver for user behavior in scenarios where a user is provided with pointers to 'related' information sources [6]. Modelling linked video in the context of serendipitous behaviour may require a completely different approach. For example, as in this case information need cannot be used as a means to limit the possible link anchors and link targets, managing the 'explosion of links' becomes a crucial part of the video linking game. However, evidence that confirms and could elaborate our model on the role of underlying information need and serendipity in video linking scenarios is lacking.

Next to the user perspective on anchor selection, an important factor in understanding user requirements with respect to linking is the user's perspective on what the links exactly should entail, the *link purpose*. For example, in the *detail-on-demand* scenario that is often referred to in the context of video linking, links are often directed towards one or a few specific information targets, such as a Wikipedia lemma, a biography or a document providing the user with background information (detail) on a selected anchor in a video. In a scenario that we refer to as *contextual linking* however, typically a plurality of links can be related to an anchor. In this scenario, the aim is to provide a user with contextual information on an anchor. For example, whereas in a detail-on-demand scenario, the anchor 'Obama' would be linked to a Wikipedia lemma, in the contextual linking scenario, the anchor would be linked to other documents that contain 'Obama'. On a global level, this may not seem to be very helpful for a user with an information need as such an approach could easily lead to a large amount of possible relevant targets. However, when the links are established

IW3C2, May 13–17, 2013, Rio de Janeiro, Brazil,
ACM 978-1-4503-2038-2/13/05.

within a closed set of documents such as a video archive or a curated collection of both internal and external information sources, the links serve as a means to structure the information available across the information sources. This would allow user to navigate through the context of an anchor via a rich link structure.

In order to validate these conceptual categorizations within the video linking framework, we need to address users that can be expected to be interested in video linking and consult them in an appropriate way. Given our experiences in the exploratory study mentioned above ([2]) that indicated that users are unfamiliar with video linking and find it difficult to reflect on it on a conceptual level, we feel that interviewing users and eliciting information based on mock-ups will only to a marginal extent provide us with the information we require.

Therefore, we propose to provide users with a scenario that approximates the technical context we are aiming for and at the same time minimize the degrees-of-freedom in the testing scenarios. To this end we are working on a series of evaluations that take the perspective of a search scenario where a particular user group has specific pre-defined information needs given a closed collection of videos. We use a video search system to perform an initial search on the video collection, ask users to assess the relevance of videos and define appropriate anchors given their information need. We use systems that participate in a benchmark evaluation on searching and linking to provide the user group with automatically selected links on the basis of both anchors that were selected earlier (known-anchor condition) and on the basis of anchors that are extracted automatically (automatic-anchoring condition). In a final step we ask users to assess the relevance of the links in the context of their information need. Following this approach we can convincingly categorise the requirements users have for linked video. In the remainder of this paper we describe the proposed methodology in more detail in Section 2. Section 3 concludes this paper.

2. REQUIREMENTS ANALYSIS

Before we describe our methodology to categorize requirements for linking, we define the components of the envisaged linking system more formally. Figure 1 shows an example of a linking system. We assume an audiovisual search system where users formulate their information need as a query and receive a ranked list of video segments as an answer. After finding a relevant video segment that fits their information need, we assume the user watches it. During this process, the system presents him/her with possible link anchors. We assume that anchors can either refer to a rectangular area in the video, a temporal sequence, or to spoken words. Note that there can be different forms of visualising the anchors. For example, if a face would be an anchor it could be highlighted. Here, we envision that the visualisation of the link anchors resides outside the display area of the video. When the user activates an anchor, typically by clicking, the system presents a textual representation of one or more link targets that it believes are appropriate to the user once followed. Following links by clicking the link target representation brings the user to the targeted information need.

The questions that this paper addresses are: first, what type of anchors users require when watching a video with respect to their information need, and second, what the characteristics of suitable link targets are. It is not the intention of this paper to provide final answers to these questions, but instead we focus on developing a methodology that will be followed by empirical validation in the future. In our previous user study [2] we found that users find it difficult to state coherent requirements when asked about an abstract technology that they have never used before. Therefore, the basic principle of our methodology is to give users the feeling that they participate in the above search and linking scenario by letting them use a state-of-the-art video search engine.

In our study we found that the considered video collection and the characteristics of the user group are important factors. Additionally, we believe that the design, functionality, and performance of the search engine plays an important role to improve the involvement of the user in the scenario. Therefore, we first describe the choices for these basic ingredients to our study before describing the methodology to answer the above questions. The actual requirements study is then divided into two parts, which we describe separately: first, we ask users to think of information needs, search for relevant video segments in the collection and mark desired anchors, second, we present the users with suggested link targets generated by several systems and ask them which ones they find appropriate.

2.1 Test-bed

We now discuss our choices of data collection, user group and search engine functionality that we will use in our requirement analysis. The video collection used for linking should at least fulfil the following requirements: the collection should be of a sufficient size to represent a real-world collection that is searched with a video search system, and it should contain content that is generally interesting to the targeted user group. We use a 200h sample of archival video as a video collection[1]. We generated the sample by randomly selecting videos from the archive so that it is representative for a collection that would be interesting for public use. The videos have a high resolution and are accompanied with manually generated archival metadata, electronic program guide (epg) information and subtitles.

Although it is clearly interesting to investigate and contrast different user groups, we focus in this study on one particular user group in order to avoid that underlying variations in use scenarios introduces noise. We choose home users as a user group as we believe that especially home users may prefer to explore collections by following links instead of entering queries. This is for example in contrast with broadcast professionals that use audiovisual search systems typically to search for reusable material and not interested in links. We employ a recruitment bureau which selected 30 users of varying age and background. To improve their intrinsic involvement, we give them a monetary compensation for their efforts.

As mentioned above, the involvement of the user also depends on the abilities of the search engine. We therefore briefly describe the capabilities of the search engine that we use. Our search engine employs several independent search services that provide confidence scores of the relevance of video segments to the current query. These confidence scores

[1]The data collection is provided by the BBC in the context of the AXES project: http://www.axes-project.eu

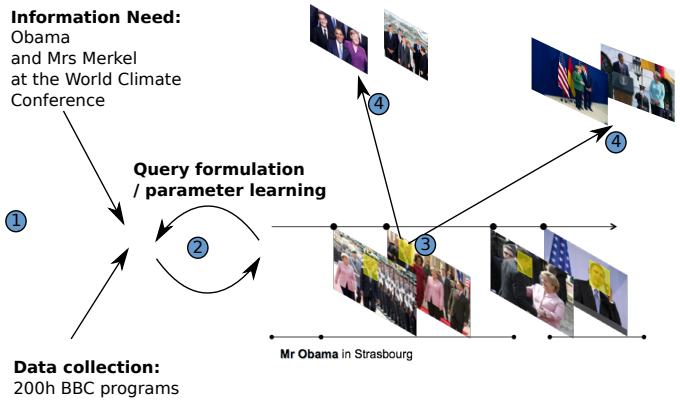

Figure 1: Conceptulization Linking

of a segment are combined by a weighted average by which the segment is ranked. We provide the users with the following search services:

Metadata service searches the metadata of a video, acting as a prior confidence for all segments in this video.

Speech service searches the occurrences of query terms in the transcripts of video segments.

Category service searches video segments that contain visual occurrences of a category, described by the user in textual form (the service downloads positive examples of this category from google prior to ranking the collection). For details please refer to [4].

Face service searches, similarly to category service, faces whose model is generated from google images for the query text. For details of the method refer to [12].

Similiarity service takes an example image and searches similar keyframes within the collection. For details please refer to [1].

Similar faces service is a specialized similarity service, which searches for images containing similar faces to the ones contained in an example image. For details please refer to [1].

The user can combine multiple services using different query terms and example images for each service. The returned confidence scores are normalized using a sigmoid function. Initially these normalized scores are equally weighted to form the final score. Upon user feedback, however, a Bayesian logistic regression function is learned to determine better weightings of the scores.

2.2 Link Anchors

We hold the first part of the study in batches of six users at a time. First, the users are given a general introduction on video search using the search engine described above. Also, the users receive an explanation on the quantity and global contents of the data collection and we give them time to explore the collection to get a general feeling of the available content. Afterwards, each user enters a text describing one or more information needs that should be satisfied by the system. Subsequently, the users are asked to search and browse the collection for video segments that are relevant to this information need and reformulate the query if needed. After a maximum of a fixed time span that will be determined during a pilot run, or after the user is convinced that additional queries will not yield additional relevant segments, we ask him/her to identify anchors in these segments. To be more concrete, for each relevant segment, we ask the user to specify the anchors that he would expect given a video segment. In order to allow the user to specify desired anchors, we allowed him/her to select rectangles inside of keyframes, spoken words, and whole segments. If the user wants a rectangle or a whole video sequence to be an anchor, we ask him/her to invent a label for this anchor segment.

After selecting the desired anchors, we gather the reasons why the user selected them. For each anchor, the user has to specify the reason for placing the anchor and an indication of what content he/she would expect if following a link behind this anchor. In particular, the user specifies whether the awaited information should provide more details on the anchor, additional information to the original information need, or whether he/she awaits serendipitous information that is unrelated to his/her original information need.

2.3 Link Targets

Given the output of the user study about possible anchors, we want to analyse requirements for the corresponding link targets. Here, a central problem is that we cannot possibly ask a user to assess all video segments in a collection and judge whether or not these are suitable targets. Instead, we could use a link generation engine to support us with suggesting possible link targets. However, using a single search engine is likely not to produce a representative set of link targets. We solve this problem by embedding the user requirement study in the MediaEval Search and Hyperlinking task 2013[2], see [5] for 2012's edition of the task. In particular, in the linking sub-task we ask participants to provide ranked lists of possible links. Here, the task input for the participants are the anchors (video sequences, rectangles or words in spoken text) defined during the first part of the requirements study. Under the assumption that linking approaches are diverse (they produce different link targets) and that most relevant links are at least returned by one participants, we use the top-ranked link targets of

[2]http://www.multimediaeval.org/

each participation as a pool of links that have to be assessed for suitability. Note that this methodology is similar to the assessment methodology in TRECVid [15]. For each link in the pool, we ask the user who defined the anchor to assess whether it is suitable or not. For each link the user also has to specify the purpose (serendipitous etc) the link would serve him/her.

2.4 User Interviews

Additionally to asking users to categorize link anchors and link targets, we also interview them after each step as to the requirement analysis. Note that during similar interviews for our previous study [2] users stated not to be able to state clear requirement. We believe that this situation will be improved now, because our users were actively involved in selected anchors and have seen link targets by current systems based on their own information needs.

3. CONCLUSIONS

We started this paper with a discussion on the lack of information on the user perspective on video linking, specifically with respect to the motivation of users to follow links, the selection of appropriate anchors for linking and the characteristics of the link targets. We described a methodology to assess these user requirements in a video-to-video linking scenario. In this methodology, a real-life scenario is mimicked using a realistic content set and a state-of-the-art audiovisual search system. We focus on a specific user group (home users) and ask them to provide information needs, search for relevant video segments given this need, indicate which elements in the video should be marked as anchors, and to assess the relevance of automatically generated link targets that are returned by participating systems in a benchmark evaluation on search and linking. We plan to discuss the first results of this evaluation during the workshop.

Acknowledgments

The work reported in this paper was funded by the EU Project AXES (FP7-269980), Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL) project at DCU; and the Dutch national program COMMIT.

4. REFERENCES

[1] R. Aly, K. McGuinness, S. Chen, N. E. O'Conner, K. Chatfield, O. M. Parkhi, R. Arandjelovic, A. Zisserman, B. Fernando, T. Tuytelaars, D. Oneata, M. Douze, J. Revaud, D. P. J. Schwenninger, H. Wang, Z. Harchaoui, J. Verbeek, and C. Schmid. AXES at TRECVid 2012. In *TREC 2012 Video Retrieval Evaluation Online Proceedings (TRECVid 2012)*, Geithesburg, U.S., December 2012. NIST.

[2] 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.

[3] M. Bron, B. Huurnink, and M. de Rijke. Linking archives using document enrichment and term selection. In *Proceedings of TPDL 2011*, pages 2357–2360, 2011.

[4] K. Chatfield and A. Zisserman. Visor: Towards on-the-fly large-scale object category retrieval. In *Asian Conference on Computer Vision*, 2012.

[5] M. Eskevich, G. J. F. Jones, M. Larson, C. Wartena, R. Aly, T. Verschoor, and R. Ordelman. Comparing retrieval effectiveness of alternative content segmentation methods for internet video search. In *10th Workshop on Content-Based Multimedia Indexing*, 2012.

[6] A. Foster and N. Ford. Serendipity and information seeking: an empirical study. *Journal of Documentation*, 59(3):321–340, 2003.

[7] A. Girgensohn, L. Wilcox, F. Shipman, and S. Bly. Designing affordances for the navigation of detail-on-demand hypervideo. In *Proceedings of AVI 2004*, pages 290–297. ACM, 2004.

[8] W. F. L. Heeren, L. B. van der Werff, R. J. F. Ordelman, A. J. van Hessen, and F. M. G. de Jong. Radio Oranje: Searching the Queen's speech(es). In C. L. A. Clarke, N. Fuhr, N. Kando, W. Kraaij, and A. P. de Vries, editors, *Proceedings of the 30th ACM SIGIR, Amsterdam*, page 903, New York, July 2007. ACM.

[9] P. Hoffmann, T. Kochems, and M. Herczeg. HyLive: Hypervideo-Authoring for Live Television. In *Changing Television Environments*, pages 51–60. Springer, 2008.

[10] B. Meixner, K. Matusik, C. Grill, and H. Kosch. Towards an easy to use authoring tool for interactive non-linear video. *Multimedia Tools and Applications*, pages 1–26, 2012.

[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 ICME 2005*, Los Alamitos, 2005.

[12] O. M. Parkhi, A. Vedaldi, and A. Zisserman. On-the-fly specific person retrieval. In *International Workshop on Image Analysis for Multimedia Interactive Services*. IEEE, 2012.

[13] I. Sawhney, N. and Balcom, D. and Smith. Authoring and navigating video in space and time. *MultiMedia, IEEE*, 4(4):30–39, 1997.

[14] F. Shipman, A. Girgensohn, and L. Wilcox. Authoring, viewing, and generating hypervideo: An overview of Hyper-Hitchcock. *ACM Trans. Multimedia Comput. Commun. Appl.*, 5(2):15:1—-15:19, 2008.

[15] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[16] S. Tan, C.-W. Ngo, H.-K. Tan, and L. Pang. Cross media hyperlinking for search topic browsing. In *Proceedings of the 19th ACM international conference on Multimedia*, MM '11, pages 243–252, New York, NY, USA, 2011. ACM.