Supporting the Interpretation of Enriched Audiovisual Sources through Temporal Content Exploration

Hugo Huurdeman, University of Amsterdam
Liliana Melgar Estrada, University of Utrecht / Netherlands Institute for Sound and Vision
Roeland Ordelman, Netherlands Institute for Sound and Vision / University of Twente
Julia Noordegraaf, University of Amsterdam

Increasingly, images and audiovisual media are being used in humanities research (Clivaz, 2016). However, moving images have been called a "blind medium" for retrieval (Sandom & Enser, 2001), since it is not possible to search their content in the same way that we search for text - necessitating the manual or automatic sequential viewing and annotation for transcoding or interpreting the audiovisual contents. This poses problems for unlocking access to large audiovisual archives in a way that they are suitable for their users, academic researchers among them (Tommasi et al., 2014).

During the Digital Humanities infrastructure project CLARIAH[1], Automated Speech Recognition (ASR) has been deployed (Ordelman et al., 2018a; Ordelman & Van Hessen, 2018) to generate speech transcripts for accessing the spoken words in audiovisual collections within the infrastructure, such as the Dutch Radio and Television collections from Netherlands Institute for Sound and Vision (NISV).[2] These automatic transcripts are available to the users of the CLARIAH Media Suite, a Virtual Research Environment (VRE).[3] The Media Suite provides access to the rich collections of audiovisual material and related contextual collections at various large heritage institutions in the Netherlands (Ordelman et al., 2018a, 2018b). It offers a Resource Viewer where each resource (e.g., a television program) can be watched together with its archival metadata (at the resource, item, and segment level), ASR transcripts, augmented by user annotation functionality. CLARIAH facilitated a research project, discussed next, to investigate how to exploit the increasing amounts of multimodal annotations from automatic text, sound and image analysis, as well as the manual annotations that scholars (or potentially also crowd-workers) can create or use via the Media Suite.

The CLARIAH ReVI project: meaningful access to temporal metadata

In this paper we present the Resource Viewer (ReVI) project, conducted as part of CLARIAH, which investigated how to support the exploration of different types of content metadata of audiovisual sources, such as segment information or automatic transcripts. Our hypothesis is that providing temporal visualizations of the content metadata might enable scholars to identify relevant items and segments, while at the same time it facilitates the

---

[1] https://clariah.nl
[2] For more information on the status of the ASR for this collection, see: https://mediasuite.clariah.nl/documentation/faq/is-data-enriched
[3] https://mediasuite.clariah.nl/

transition between distant and close reading (van der Molen, Buitinck & Pieters, 2017) and a more 'scalable' approach to "connecting the distant with the close" (Denbo and Fraistat 2011).

The created ReVI Timeline prototype (see Figure 1) offers a temporal display of the content metadata (e.g., of the ASR), *temporal tag clouds*, and a layered approach[4] to presenting and creating transcripts and annotations. Tag or word clouds have proved to be useful in the textual domain in providing overviews (Heimerl et al., 2014) or, in a media studies context, for detecting "busts of attention" (van der Molen & Pieters, 2017). We thus hypothesize that incorporating a temporal dimension to word clouds based on the time-based media's content metadata could provide similar benefits. This has rarely been done in previous literature and interfaces[5]. In our case, temporal word clouds were created by dividing an ASR transcript into a number of predefined segments, and extracting the salient words (or combinations of words) for each segment using the TF-IDF algorithm[6] and Part-of-Speech tagging[7].

To evaluate the Timeline Prototype in general, and the temporal tag clouds in particular, we conducted a formative usability study[8] (Baxter, 2015). This study was performed in a lab setting using a think-aloud protocol, utilizing NISV collections with ASR transcripts. Five scholars with a Media Studies background (television studies, film studies & new media) and experience in using the Media Suite participated. Previous studies (e.g., Bron et al, 2016; Melgar Estrada et al, 2017) have identified stages in scholarly work with digital collections: an exploration and assembling stage of research (when scholars try to collect relevant sources, i.e., their "corpus"), and a focused analysis stage of research (when scholars close read their sources and analyze them). Accordingly, two simulated work tasks (Borlund, 2003) were used to look at the use and usefulness of temporal content visualizations with respect to these two different stages in the research process of media scholars.

Discussion

The results of our user study showed the added value of visualizing the content of audiovisual media via the temporal aspects described by archival metadata, ASR, and manual annotations. Participants indicated feeling "increased agency" with the use of overviews provided by the temporal tag clouds, together with the option to search and highlight keywords, and zooming in and out to specific segments based on their queries. With respect to the usefulness of these tag clouds in the research process, participants in the study indicated that the temporal tag clouds decline in value over time, after initial exploration of a resource. When a researcher builds up a detailed knowledge frame of a resource, automatic

---

[4] A tier-based data model is used by multimedia annotation tools for multimodality research such as ELAN (Sloetjes, 2014).

[5] In another context, Jatowt et al (2011), depicted web page evolution over time using word clouds.

[6] Following Jatowt et al (2011)'s definition of salient terms, "Salient terms in a given, target time segment T_w are terms that have high scores inside T_w and, at the same time, have low scores inside other time periods."

[7] Future work includes the analysis of other possible models, e.g., parsimonious language models, for the generation of the word clouds (Kaptein et al., 2010).

[8] Formative studies are commonly "done early in the product development life cycle to discover insights and shape the design direction" (Baxter, 2015)

tag clouds might lose their added value (Huurdeman et al, 2016). We assume that this happens since once the "corpus" is selected, scholars go into a more defined and personalized analysis of their sources based on their research questions, using their own categorizations.

Based on the users' feedback, various challenges came up which underline the importance of a data and tool criticism approach (Koolen, van Gorp & Ossenbruggen, 2018) to the Resource Viewer as part of a digital research environment. First, there is the issue of maintaining and displaying provenance in terms of data, given the abundance of metadata created in different ways. Second, inherent quality issues of the automatic enrichments might occur[9]. Third, transparency in terms of visualization is important. Automatically generated tag clouds, by juxtaposing temporally occurring words, might show unrelated words side-by-side, thus potentially suggesting links that are not present in the content. In our study, we saw instances of "positive serendipity" -- words sparking genuine ideas, versus "negative serendipity" -- false impressions due to visualization issues. To address these issues, we revised the Timeline Prototype, adding a way to view the source data and data transformations of the automatic tag clouds, adapting their layout, and adding visual indications of potentially unreliable ASR sentences.

Conclusion and future work

Ample challenges exist for building tools that provide meaningful ways to make sense of the increasing amount of automatically and manually generated metadata outside the traditional cycles of archival curation.

The Timeline Prototype created in the ReVI project showed how temporal visualizations and advanced annotation features potentially facilitate looking beyond the picture and into the temporal aspects of audiovisual content. This type of content exploration can benefit scholars doing research with audiovisual sources (e.g., in media studies, oral history, film studies, and other disciplines which are increasingly using AV media). The participants' feedback showed that transparency of tools and respecting provenance information about the data is of utmost importance. The final outcomes of the ReVI project can serve as an inspiration for improving AV-media-based research tools[10], in particular in the phase of exploratory search for corpus building.

References

Baxter, Kathy. Understanding Your Users (Interactive Technologies). Elsevier Science. Kindle Edition.

[9] For instance, since the ASR algorithm was trained to recognize Dutch language, there is resulting noise when non-Dutch languages are used in an audiovisual item.

[10] For instance, in autumn 2019, elements of the ReVI prototype will start to be integrated into the CLARIAH Media Suite; future evaluations and workshops are expected after that.

Borlund, P. (2003). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. Information Research, 8(3). Retrieved from http://www.informationr.net/ir/8-3/paper152.html

Bron, M., van Gorp, J., & de Rijke, M. (2016). Media studies research in the data-driven age: How research questions evolve. Journal of the Association for Information Science and Technology, 67(7), 1535–1554. https://doi.org/10.1002/asi.23458

Clivaz, C. (2016, June). *Keynote address*. Presented at the AV in DH workshop at the Digital Humanities Conference, Krakow, Poland. Retrieved from https://avindhsig.wordpress.com/workshop-2016-krakow/accepted-abstracts/keynote-address-dr-claire-clivaz/

Denbo, S., and Fraistat, N. (2011). "Diggable Data, Scalable Reading and New Humanities Scholarship". In 2011 Second International Conference on Culture and Computing, 169–70. https://doi.org/10.1109/Culture-Computing.2011.49

Huurdeman, H. C., Wilson, M. L., & Kamps, J. (2016). Active and Passive Utility of Search Interface Features in Different Information Seeking Task Stages. Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval, 3–12. https://doi.org/10.1145/2854946.2854957

Jatowt, A., Kawai, Y., & Tanaka, K. (2011). Page History Explorer: Visualizing and Comparing Page Histories. IEICE TRANSACTIONS on Information and Systems, 94(3). Retrieved from http://www.dl.kuis.kyoto-u.ac.jp/~adam/ieice11.pdf

Kaptein, R., Hiemstra, D., & Kamps, J. (2010). How Different Are Language Models and Word Clouds? In C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, … K. van Rijsbergen (Eds.), *Advances in Information Retrieval* (pp. 556–568). Springer Berlin Heidelberg.

Koolen, M., van Gorp, J., & van Ossenbruggen, J. (2018). Toward a model for digital tool criticism: Reflection as integrative practice. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqy048

Melgar, L., Koolen, M., Huurdeman, H., & Blom, J. (2017). A Process Model of Scholarly Media Annotation. In Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (pp. 305–308). New York, NY, USA: ACM. https://doi.org/10.1145/3020165.3022139

Ordelman, R. J. F., & Hessen, A. J. van. (2018). Speech Recognition and Scholarly Research: Usability and Sustainability. *CLARIN 2018 Annual Conference*, 163–168. Retrieved from https://research.utwente.nl/en/publications/speech-recognition-and-scholarly-research-usability-and-sustainab

Ordelman, R., Melgar Estrada, L., Martínez Ortiz, C., Noordegraaf, J., & Blom, J. (2018a). Media Suite: Unlocking Archives for Mixed Media Scholarly Research. In CLARIN Annual Conference. Pisa, Italy. Retrieved from https://www.clarin.eu/clarin-annual-conference-2018-abstracts

Ordelman, R., Martínez Ortíz, C., Melgar Estrada, L., Koolen, M., Blom, J., Melder, W., de Boer, V., Karavellas, T., Aroyo, L., Poell, T., Karrouche, N., Baaren, E., Wassenaar, J., Noordegraaf, J., Inel, O. (2018b). Challenges in Enabling Mixed Media Scholarly Research with Multi-media Data in a Sustainable Infrastructure. In Digital Humanities Conference. Mexico: ADHO. Retrieved from https://dh2018.adho.org/en/challenges-in-enabling-mixed-media-scholarly-research-with-multi-media-data-in-a-sustainable-infrastructure/

Sandom, C., & Enser, P. G. B. (2001). VIRAMI: visual information retrieval for archival moving imagery. Presented at the International Cultural Heritage Informatics Meeting, Milano, Italy: Archives & Museum Informatics. Retrieved from http://www.archimuse.com/publishing/ichim01_vol1/sandom.pdf

Sloetjes, H. (2014). ELAN. *The Oxford Handbook of Corpus Phonology*. https://doi.org/10.1093/oxfordhb/9780199571932.013.019

Tommasi, T., Aly, R., McGuinness, K., Chatfield, K., Arandjelovic, R., Parkhi, O., … Tuytelaars, T. (2014). Beyond metadata: searching your archive based on its audio-visual content. *Proceedings of the 2014 International Broadcasting Convention*. Presented at the IBC, Amsterdam. https://doi.org/10.1049/ib.2014.0003

van der Molen, B., & Pieters, T. (2017). Distant and Close Reading of Dutch Drug Debates in Historical Newspapers: Possibilities and Challenges of Big Data Analysis in Historical Public Debate Research. In A. K. Somani & G. C. Deka (Eds.), *Big Data Analytics: Tools and Technology for Effective Planning.* (pp. 373–390). Retrieved from http://public.eblib.com/choice/publicfullrecord.aspx?p=5116734

van der Molen, B., Buitinck, L., & Pieters, T. (2017). The leveled approach. Using and evaluating text mining tools AVResearcherXL and Texcavator for historical research on public perceptions of drugs. ArXiv:1701.00487 [Cs]. Retrieved from http://arxiv.org/abs/1701.00487

Appendix

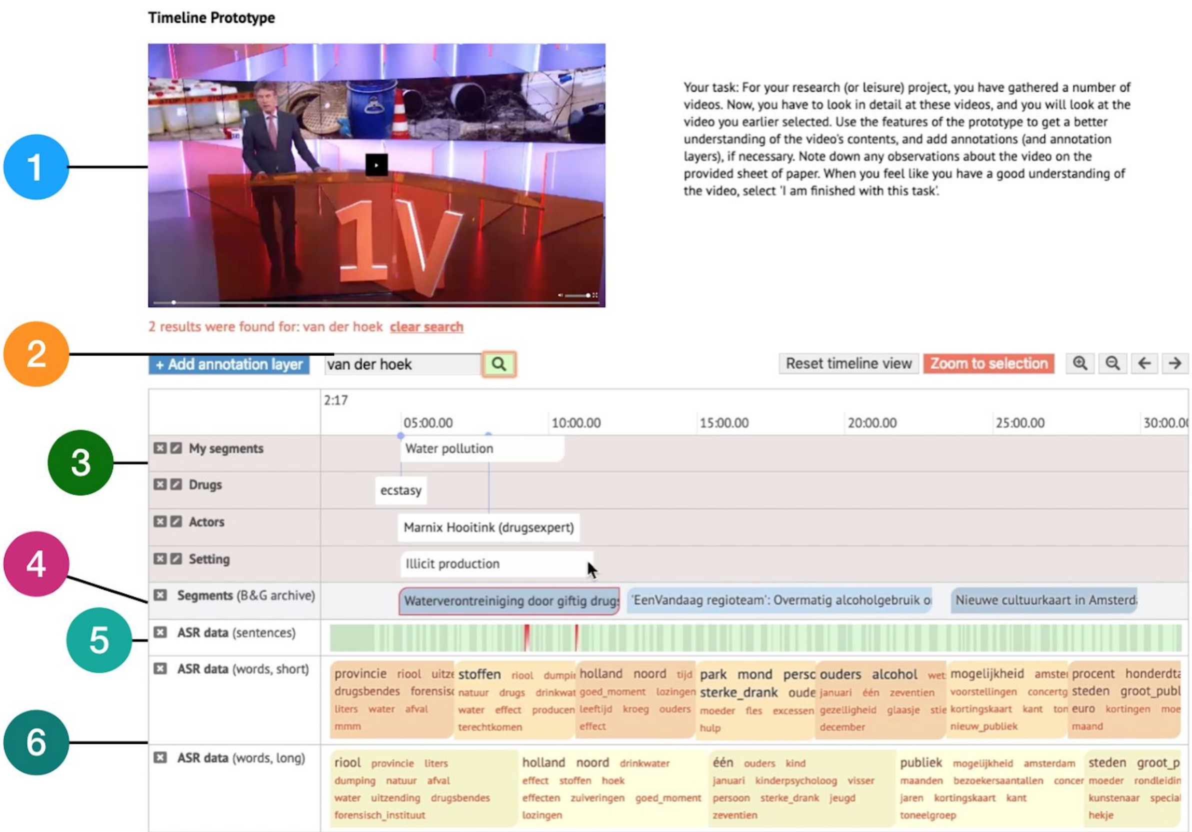

Figure 1. Timeline Prototype evaluated in user study. (1) video player, (2) ASR search functionality (3) multi-tiered annotation functionality, (4) visualization of metadata-based segments, (5) visualization of ASR at sentence level, and (6) automatic ASR temporal tag clouds

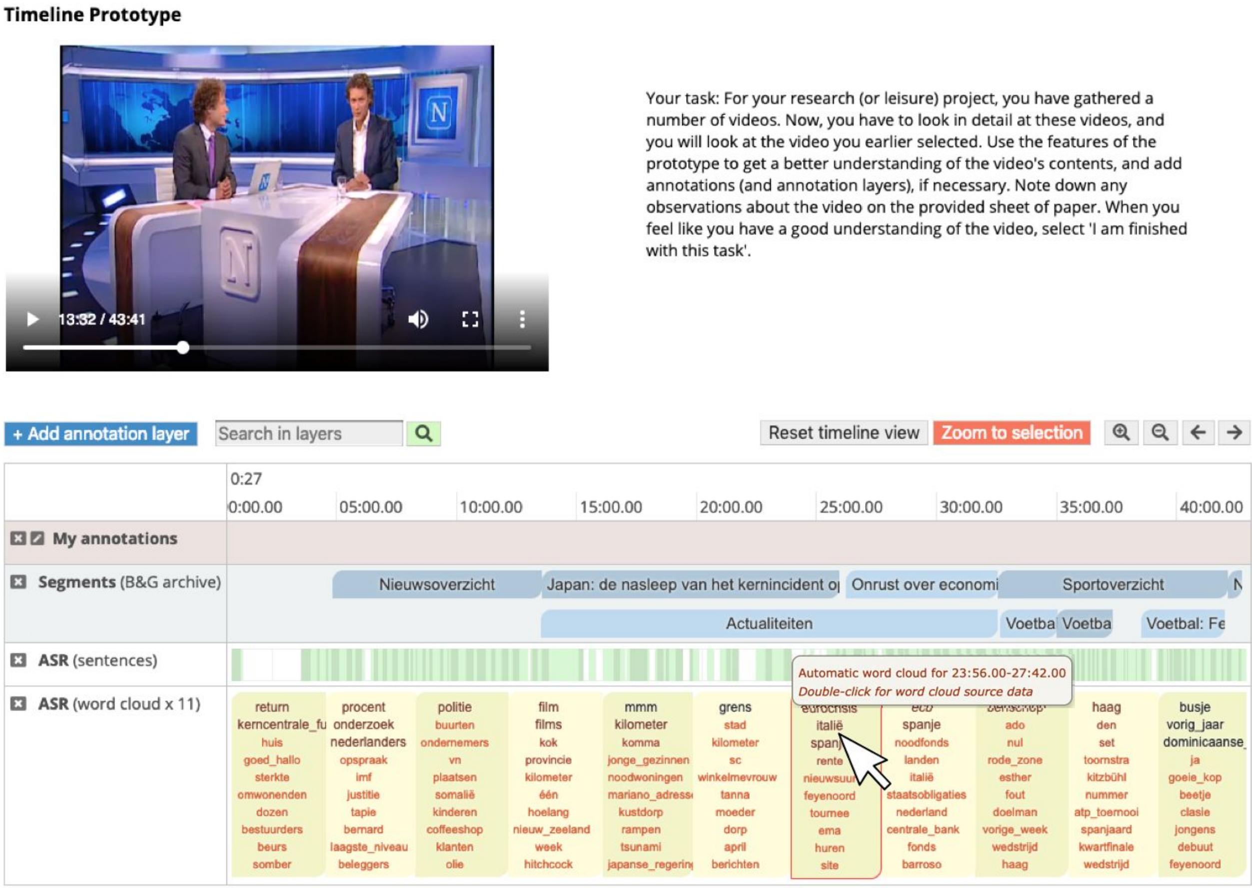

Figure 2: Revised prototype. ASR reliability has been added using color coding, the formatting of temporal tag clouds has been adapted, and the possibility of viewing source data added