Automatic Annotations and Enrichments for Audiovisual Archives

Nanne Van Noord[1,2], Christian Gosvig Olesen[3], Roeland Ordelman[2,4] and Julia Noordegraaf[1]
[1]*University of Amsterdam, The Netherlands*
[2]*Netherlands Institute for Sound and Vision, The Netherlands*
[3]*Utrecht University, The Netherlands*
[4]*University of Twente, The Netherlands*

Keywords: Audiovisual Archives, Automatic Speech Recognition, Computer Vision.

Abstract: The practical availability of Audiovisual Processing tools to media scholars and heritage institutions remains limited, despite all the technical advancements of recent years. In this article we present the approach chosen in the CLARIAH project to increase this availability, we discuss the challenges encountered, and introduce the technical solutions we are implementing. Through three use cases focused on the enrichment of AV archives, Pose Analysis, and Automatic Speech Recognition, we demonstrate the potential and breadth of using Audiovisual Processing for archives and Digital Humanities research.

1 INTRODUCTION

Despite the massive progress made in the performance levels and overall availability of Audiovisual Processing (AVP) technologies, such as Computer Vision (CV) and Automatic Speech Recognition (ASR), the availability of these tools for heritage institutions is still lagging behind. As a consequence, Humanities researchers, in particular Media Studies scholars, cannot fully benefit from the large amounts of digitized and digital born heritage data that have been successfully made available in the past decades in the course of large digitization programs and initiatives via (research) data infrastructures. The fact that AVP tools are underexposed in Digital Humanities (DH) is partly due to the predominance of text-centric scholarship (Wevers and Smits, 2020). However, for the analysis of audiovisual resources (including both video materials and images) textual representations (e.g., transcripts, semantic annotations) are often sparse or nonexistent. Moreover, AVP and DH have different analytical foci and ambitions and may, as a consequence, have difficulties in productively combining perspectives into a shared research agenda.

In recent years there have been several efforts to bridge the gap between AVP, Media Studies, and DH. For instance, the recent Sensory Moving Image Archive (SEMIA) project[1] brought together researchers from Media Studies and Computer Science, with design practitioners, and experts from AV archives to develop novel ways of exploring AV archives based on syntactic features. Syntactic features concern the visual appearance of the material, without resorting to semantic categorisations (e.g., colour or shape descriptors), which presents a paradigm shift as compared to traditional metadata but nonetheless connects well to existing and ongoing research efforts. Similarly, the I-Media-Cities project[2] – a collaboration between 9 European film heritage institutions – developed a search interface offering search and video annotation tools that allow researchers to easily retrieve and categorize shots based on various semantic and syntactic categories. Besides such projects, DH scholars have explored the affordances of various open source visual analysis software suites - for instance ImageJ - for stylometric research, as a means to study patterns in the works of specific directors, genres, production companies, painters or social media images (Manovich, 2013). These works offer interesting points of comparison because of their efforts to use AVP for DH research. Yet, the results of these efforts have not led to an increase in practical availability of AVP and their applicability remains limited to a very small set of features and types of collections (Kuhn et al., 2014).

A major challenge for making AVP tools available is that the underlying algorithms need direct access to

[1]https://sensorymovingimagearchive.humanities.uva.nl/

[2]https://www.imediacities.eu/

Van Noord, N., Olesen, C., Ordelman, R. and Noordegraaf, J.
Automatic Annotations and Enrichments for Audiovisual Archives.
DOI: 10.5220/0010387706330640
In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART 2021) - Volume 1*, pages 633-640
ISBN: 978-989-758-484-8

the data, yet due to strict IPR regulation this data often cannot leave the archives' digital environment. A side effect of this is that for the analysis of AV material commercially available tools cannot be used, as these are primarily cloud based. To tackle this challenge it is necessary that any AVP infrastructure can be integrated with a local archival infrastructure, i.e., we need to bring the algorithms to the data, rather than the data to the algorithms. This approach is incorporated in the design of the Common Lab Research Infrastructure for the Arts and Humanities (CLARIAH), a DH research infrastructure. However, approaching the data access challenge in this manner creates a new challenge, namely the availability of usable AVP tools. Although in recent years we have seen an increase in the publishing of open-source code in AVP, this code is often only geared towards reproducing the results in the accompanying paper. Building a tool that can be applied to novel datasets on top of this code can be non-trivial. Dealing with this latter challenge is one of the foci of the CLARIAH PLUS project. In this paper we describe the steps taken so far and the future plans to overcome this challenge and to make AVP tools practically available for use by DH scholars.

In the following we will give a brief introduction of AVP and the CLARIAH AVP infrastructure in Section 2. In Section 3 we discuss three use cases that demonstrate the potential of AVP for archives and media research. In Section 4 we summarise our progress until now and reflect on future steps.

2 AUDIOVISUAL PROCESSING AND THE DIGITAL HUMANITIES

The unlocking of audiovisual collections with AVP approaches, while still limited in scope, has so far proven valuable to DH researchers with an interest in media history and social history (Arnold and Tilton, 2019; Wevers and Smits, 2020; Bhargav et al., 2019; Lincoln et al., 2020). Similarly, among AVP researchers the benefits of collaborative efforts with DH researchers for analyzing visual, cultural resources has also been put forward. For example, Ahmed Elgammel, director of the Art & AI lab at Rutgers University, made such a case[3] when he argued in 2014 that visual art is an excellent testbed for studying and replicating human perception in computers. Which they demonstrate with a method for automatically discovering artistic influence in a large database of paintings (Saleh et al., 2016). Yet, despite a shared desire for more collaboration, and various initiatives to bridge the gaps between the fields, the practical availability of AVP tools and progress on shared research agendas remains limited.

This lack of shared research agenda's can partly be explained by a difference in focus. In the field of DH, the application of AVP tools to large image datasets significantly scales up the level of analysis from 'close' to 'distant' viewing (Arnold and Tilton, 2019), allowing scholars to observe patterns in the data that are invisible in the traditional, interpretative analyses of small sample sets. Whereas such usage, thus, is a significant innovation in the field of the humanities, in the field of AVP research the research agenda is primarily driven by technical innovation of the tooling itself, as a consequence of which such application-oriented research in a specific domain is not considered in scope. Conversely, from a traditional humanities perspective the types of analyses performed by AVP researchers are considered much more shallow on a conceptual level (for instance in terms of historiography or theories of visual culture) than the complex analyses of meaning that the interpretative, hermeneutic method of close reading allows for. Thus, what might be revolutionary in the field of AVP often has little to no impact in the humanities, as it is simply a reproduction of known metadata or grounded in outdated theories and views. This is exemplified by the criticism of Griselda Pollock[4], Professor of Art History on the method for discovering artistic influences by Saleh et al. (2016), likening it to the work of 19th century art historians and noting that it can only lead to "*superficial resemblances at which any artist would laugh*".

The analysis of visual resources with digital tools is equally affected by the gap between AVP and DH. Currently, digital tools for paper resources such as magazines and periodicals are predominantly text-centric, and have tended not to consider the strong presence of visual elements in them. Due to the lack of practically available AVP tools to digital humanities, scholarly research has '[...] grossly neglected visual content, causing a lopsided representation of all sorts of digitized archives' (Wevers and Smits, 2020). In response to such observations a number of collaborations between AVP toolmakers and DH scholars have emerged, such as the Distant Viewing Toolkit (Arnold and Tilton, 2019), as well as the efforts described in this paper. A key focus of these collabora-

[3]https://theconversation.com/computer-science-can-only-help-not-hurt-art-historians-33780

[4]https://theconversation.com/computers-can-find-similarities-between-paintings-but-art-history-is-about-so-much-more-30752

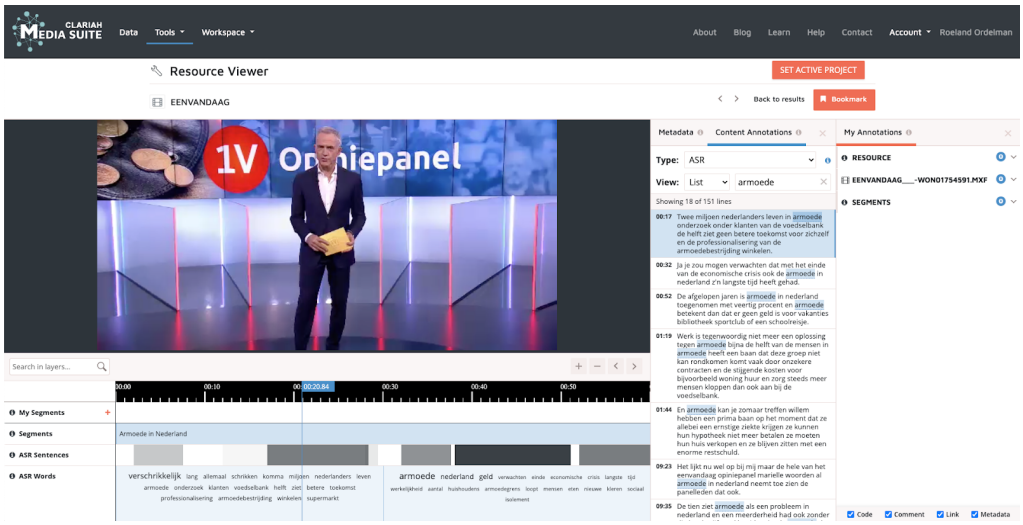

Figure 1: Screenshot of the 'resource viewer' of the Media Suite that enables browsing via time-coded speech transcripts.

tions is to take state of the art algorithms and make them available as open source solutions, making it possible for DH scholars to use them, and AVP researchers to build on them.

2.1 CLARIAH and the Media Suite

CLARIAH is a distributed data & tools infrastructure for the Humanities and Social Sciences, funded by the Netherlands Organization for Scientific Research in two subsequent projects, respectively CLARIAH-CORE (2014-2018) and CLARIAH-PLUS (2019-2024). In CLARIAH-CORE basic infrastructure and tools were developed following user-centered design principles and by carrying out research pilots so as to develop a detailed understanding of researchers' needs. Building on the experiences and insights from CLARIAH-CORE, CLARIAH-PLUS simultaneously focuses on scaling up and facilitating the implementation of the CLARIAH infrastructure in research and teaching. CLARIAH's data and tools are stored at the partner institutions (so called CLARIAH centers: knowledge centers, libraries, archives and museums) that make the large data collections accessible via online virtual research environments (VREs) for users to access on their home computers.

An example of such a VRE is the Media Suite[5] (Melgar-Estrada et al., 2019; Ordelman et al., 2019), developed by a multidisciplinary team, including media scholars, computer scientists, programmers and staff at the partner institutions. The Media Suite offers access to a highly diverse range of collections, including films, TV, radio, newspapers, oral history interviews, posters, and historical business documents, preserved and digitized by among others The Netherlands Institute for Sound and Vision (NISV), Eye Filmmuseum and the National Library of the Netherlands. With regard to the collections of NISV and Eye, it is the first time that these two collections are opened for online viewing to a scholarly audience (with a university login) outside of the institutions.

As part of the endeavour to make these collections available in the Media Suite in CLARIAH-CORE, it was a key ambition to unlock the collections embedded by adding data enrichments in addition to institutional metadata. For instance, in the context of the Media Suite pilot "MIMEHIST: Annotating Eye's Jean Desmet Collection" (2017-2018), we used a combination of various approaches for the analysis of text and images so as to facilitate browsing of historical business documents from the collections as visual as much as text resources (Olesen and Kisjes, 2018). This produced new classifications of the business archive of cinema owner and film distributor Jean Desmet (1875-1956) that complemented the Eye Filmmuseum's current finding aid while producing new data for the Media Suite. Ultimately such enrichments may support researchers in establishing links between films and contextual documents, for instance to explore recurrent (visual) motifs in both. Experimental in nature, this project was in part a precursor to the DANE environment and its ambition to deploy AVP tools in the Media Suite, that we discuss in more detail in the sections below.

[5]https://mediasuite.clariah.nl/

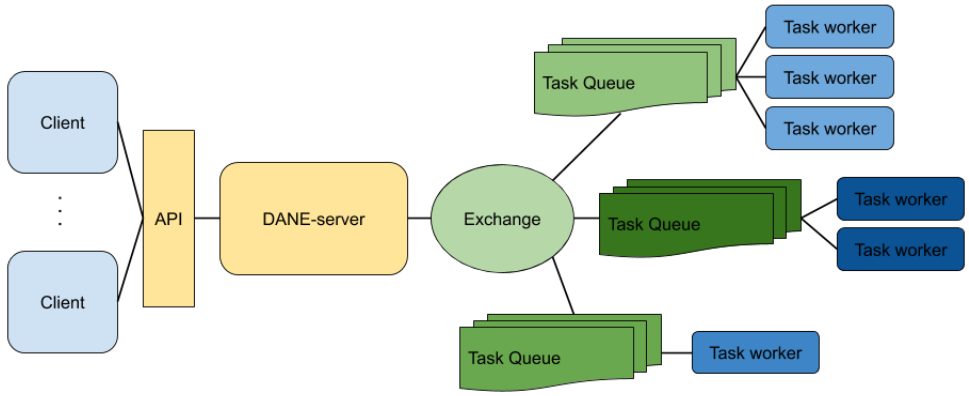

Figure 2: Schematic overview of the DANE architecture that illustrates the components that mediate the interaction between clients (the consumers of the results), and the workers (the producers of the results).

2.2 CLARIAH AVP Infrastructure

As part of the CLARIAH infrastructure we have developed a processing environment that is optimised for deploying AVP tools efficiently on (high performance) computer clusters in a transparent and reproducible way. This environment, called DANE (Distributed Annotation 'n' Enrichment)[6] , provides the framework needed for processing AV data and is designed to allow researchers maximum flexibility when deploying AVP algorithms. In addition, it keeps track of and provides researchers with clear insight into the operations that are performed on the data, ensuring that research can perform data and tool criticism at all stages of the research process.

A core design principle for DANE was to make the conversion process of an algorithm (or any piece of code) to a worker (a component which performs the data analysis) as effortless as possible. The aim here was to encourage uptake by algorithm developers, ensuring that state of the art of the algorithms continue to be made available. Moreover, DANE is designed to be able to function in environments where there are fewer compute resources, or environments where access to the data is rate limited, as is typical for AV archives. To make this possible, the DANE architecture (as shown in Figure 2) divides the tasks to be performed over different workers that perform one specific type of analysis, where the work is assigned to a worker by means of message broker queuing system. The results of the analyses are communicated back to the server by the workers, and made available via an API to clients (e.g., other services or individual users).

[6]https://github.com/CLARIAH/DANE

3 USE CASES

In this section we will outline three use cases where AVP tools are employed to enrich AV material, starting with a use case that highlights the breadth of AVP, and concluding with two use cases that focus on visual and audio analysis, respectively. Additionally, these use cases also demonstrate the different levels of technical complexity, ranging from tools intended to be integrated in the Media Suite, to completely custom and novel analysis, to tailored analyses that align with familiar methodologies.

3.1 Enriching AV Archives

Interfaces for AV archives, such as the Media Suite, center on metadata. In addition to enabling search, browsing, or viewing a single item, the metadata is crucial for contextualising the AV material and thus for the way its content may be interpreted. The majority of the metadata in AV archives is about the material, and its production, rather than about the content. While most material has descriptors such as a title and a synopsis, there is little to no metadata about the visible or audible aspects of the material. As a result it can be nearly impossible to find results for certain queries in AV archives. For example, a researcher might want to document the public debate concerning refugees, and relate this to visual depictions of refugees. Without enriched metadata the only way to tackle such a project is to watch many hundreds of hours of news and current affairs programs.

To enrich the AV collections in the Media Suite with automatically generated (and time-coded) metadata an assortment of AVP tools had been added to DANE, ranging from Semantic Object Detection, to colour analysis, to Automatic Speech Recognition.

These tools generate two types of metadata: (i) structured data which can be integrated into the existing metadata systems, and (ii) feature vectors which describe the data on a more abstract level, but which can be integrated in new types of interfaces and search systems. For instance, with an Object Recognition model it is possible to recognise which objects are present in an image, but the output of such a model can also be used to perform more abstract visual similarity queries (Smeulders et al., 2000; Babenko and Lempitsky, 2016). This latter type of query enables the modes of exploration which were central to the SEMIA project, potentially even returning results with different semantics, but which are similar on a visual syntactic level. As part of this use case we aim to integrate both types of results into the Media Suite, on the one hand enabling text-based queries for visual concepts (e.g., for detected object classes), and on the other hand resolving visual similarity queries using the extracted feature vectors.

In addition to adding benefits to the Media Suite itself, automatically enriched metadata can also enable and empower downstream applications. By making available the enriched metadata through an API it will become possible for third-party developers and researchers to find novel and innovative uses for the data. For instance, in recent years we have seen an increased uptake of the Generous Interface principles proposed by Mitchel Whitelaw (2015), in interfaces for exploring digital heritage collections. Building on the concept of the Information Flaneur (Dörk et al., 2011), Whitelaw argues that interfaces for Cultural Heritage collections should move away from keyword search and instead provide "rich, navigable representations of large digital collections; [that] invite exploration and support browsing, using overviews to establish context and maintain orientation while revealing detail at multiple scales." The enriched metadata extracted with AVP algorithms lends itself incredibly well for Generous Interfaces, moreover especially for the extracted feature vectors such interfaces may indeed be the only manner to incorporate them into an interface.

A second downstream application that benefits from AVP enriched metadata are Data Stories, a practice that emerged in the fields of Data Journalism and Narrative Visualisation (Segel and Heer, 2010) where written stores are combined with (interactive) data visualisations. A data story about the popular Dutch chat show DWDD[7] demonstrates the potential of using AVP enriched metadata. Combining Automatic Speech Recognition, Image recognition, and traditional metadata, this data story gives an overview of the TV show while also revealing peculiarities which might provide entry points for closer investigation.

While these examples highlight some of the benefits of incorporating automatically enriched metadata into AV archives, their potential has not yet been fully explored. Moreover, there are still many unanswered questions about optimal archiving procedures and best practices for this type of data. However, we hope that as enriched metadata become more established and embraced by the wider archiving community these challenges will be tackled as well.

3.2 Pose Analysis for Genre Classification

Early cinema is heavily influenced by traditions and practices from theatre (Szeto, 2014). Among others, this influence can be observed in the movements and poses of actors. Because subtle poses do not come across on a large stage, theatrically trained actors were used to exaggerating their movements to reach the audience. When actors and directors transitioned from theatre to film such practices transitioned along with them. In this use case we are investigating the use of pose detection algorithms to characterise movements in early cinema, to determine whether it is possible to discern developments in film genres and conventions of movement based on the poses of the authors. For instance, a common assumption is that Scandinavian, German and Italian dramas developed distinct acting styles that could be considered more cinematic in their use of cinematic space, while the acting styles of farces and comedies were considered as pertaining to popular, theatrical genres such as vaudeville.

Methodologically, current film historical research on film acting's development in relation to cinematic rhythm, editing, and space has relied heavily on manual annotation approaches and human reenactment based on the Laban movement analysis (LMA) method in combination with video annotation software (Pearlman, 2009; Oyallon-Kololski and Williams, 2018; Ruszev, 2018). As such research requires fine-grained and labour-intensive annotation, they often remain limited to single films, or smaller corpora. This use case offers new entry points for such research by allowing for exploring and browsing poses in a larger corpus, such as the Desmet Collection; a diverse corpus consisting of multiple genres of early cinema from various countries, that is available through the Media Suite.

To explore the relationship between pose and genre we rely on a mixture of supervised and unsupervised techniques to determine whether pose infor-

[7] https://mediasuitedatastories.clariah.nl/DWDD/

Figure 3: Visualisation of keypoint detection on a still from the Desmet collection. The coloured boxes indicate detected persons, and the detected keypoints are indicated by dots and connected to other keypoints with coloured lines.

mation can be used as a predictor of genre, but also to see whether clustering pose information would lead to a natural separation between genres. To describe pose information we extend the pose representation of (Jammalamadaka et al., 2012) to include full body, rather than only upper body keypoints. Keypoints are annotations of specific points on a human body, typically focusing on joints or extremities (e.g., top of the head, nose, elbow, hip, foot), that make it possible to describe the pose that the body is in. For this project we use the automatic keypoint detector of the Detectron2 framework (Wu et al., 2019). This detector is able to recognise up to 17 keypoints which makes it possible to capture a wide variety of poses. While the project is still ongoing, the initial analyses (as illustrated in Figure 3) show that this keypoint detection approach is sufficiently robust for applying it to historical material, highlighting the feasibility of using AVP techniques for historical film analysis.

3.3 Speech Recognition for Interview Collections

Since many years, Oral Historians have been promised the virtues of automatic speech recognition (ASR) in support of two important scholarly primitives (Unsworth, 2000; Blanke and Hedges, 2013): (i) *annotation*, creating verbatim transcripts of spoken data (such as interviews) with the use of speech recognition, and (ii) *discovery*, searching audiovisual collections using indexed speech transcripts (Gustman et al., 2002; de Jong et al., 2008). In the past decades, we have been working with Oral Historians on topics related to speech recognition in a variety of projects such as CHoral (Heeren et al., 2009), Verteld Verleden (Ordelman and de Jong, 2011), and Oral History Today (Kemman et al., 2013). More recently, besides Oral Historians also scholars in other fields that work with interview collections (e.g., with patients in the medical domain) have expressed interest in incorporating ASR in their methodology.

As a support tool for Oral Historians or more general, scholars working with interview data, the use of ASR should fit into the transcription and annotation workflows that scholars are used to, including the tools they use for analysis of the data. Also, ASR should at least be able to provide *approximately correct* transcriptions, and should facilitate error correction when the quality is (partly) insufficient (e.g., with respect to proper names or locations that are 'out-of-vocabulary'). Out-of-vocabulary (OOV) refers to the fact that speech recognition systems need to "know" a word (have a word in its vocabulary) before it can be recognized.

Hence, the quality of the transcriptions is an important requirement for scholars using automatic speech recognition. Quality typically translates to word error rate (WER): the number of errors divided by the total number of words spoken. As the multisemiotic nature of audiovisual data adds dimensions for inquiry that do not exist in written text (Goldman et al., 2005), scholars working with AV are used to create manual transcriptions and annotations based on their codebooks. The creation of accurate manual transcriptions that account for all speech events as well as other metadata (such as speaker identities and changes) or codes from a scholar's codebook (e.g., based on intonation of a speaker's voice), can take up to 50 times real-time, depending on the nature of the data and the level of detail (Barras et al., 2001). Support from a tool such as ASR (but also speaker recognition) could therefore be very helpful, especially as the technology also provides time-labels that link the transcript directly to the locations in the data.

We know from research in the field of Information Retrieval that there is a near-linear relationship between quality of transcriptions and search performance (Garofolo et al., 2000), but also, that even if transcriptions are *approximately correct* such transcriptions provide a useful basis for searching. However, scholarly discovery may in cases depend on a speech recognition system *not* making a notorious type of error: substituting a spoken word for another word as the spoken word is OOV. Typically this happens when applying speech recognition trained on a general domain to a specific topic: typically topics that could be discussed during an Oral History interview. Especially names and locations, but also content words related to a special topic, will be OOV and will consequently not be recognized, and in turn, never be found during searching.

Making automatic speech recognition available for Oral Historians and other Humanities scholars in a research infrastructure setting (bringing the algorithms to the data) to foster the analysis of audiovisual resources, requires a deliberate technical design given the above mentioned requirements. The design of the DANE environment discussed in section 2.2 warrants measures to adapt to these requirements: secure access to and processing of audiovisual sources of institutional collections in bulk, efficient use of dedicated machinery (e.g., local or cloud-based computer clusters), dealing with dynamics with respect to robustness, latency, process management, and storage of intermediate data, the implementation of adaptive workflows to address the OOV problem, and keeping track of provenance information (which version of a speech recognition system was used).

4 CONCLUSION

In this article we have argued that there is a gap between what is possible in terms of Audio Visual Processing and what is available to DH scholars, and that to bridge the gap between these fields it is necessary to make available established tools. On the one hand, this will enable DH scholars to incorporate AVP approaches and technologies in their research, to gain a fuller, more 'macroscopic' perspective on audiovisual media (Graham et al., 2015); on the other, the semantically complex analyses of DH scholars can be used as input to boost the development of semantically sensitive AVP algorithms.

While a major step in this process is to make AVP tools practically available, it is equally necessary to create an environment in which tools can be developed or customised to support answering newly emerging research questions. By integrating DANE, our proposed environment for deploying AVP tools, in the Media Suite virtual research environment we have been able to bring the algorithms to the rich catalogue of datasets that is available in the Media Suite. Moreover, the distributed and modular design of DANE ensures flexibility in deploying new tools, as well as an easy and well-documented process for converting algorithms to tools. In embracing such an approach we have taken the first steps in developing an AVP environment within CLARIAH that enables a continuous cycle of automatically annotating and enriching AV archives, opening the door for further collaboration between AVP researchers and DH scholars.

ACKNOWLEDGEMENTS

The research described in this paper was made possible by the CLARIAH-PLUS project (www.clariah.nl) financed by NWO.

REFERENCES

Arnold, T. and Tilton, L. (2019). Distant viewing: Analyzing large visual corpora. *Digital Scholarship in the Humanities*.

Babenko, A. and Lempitsky, V. (2016). Efficient Indexing of Billion-Scale Datasets of Deep Descriptors. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2055–2063, Las Vegas, NV, USA. IEEE.

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (2001). Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1-2):5–22.

Bhargav, S., van Noord, N., and Kamps, J. (2019). Deep learning as a tool for early cinema analysis. In *Proceedings of the 1st Workshop on Structuring and Understanding of Multimedia heritAge Contents*, pages 61–68.

Blanke, T. and Hedges, M. (2013). Scholarly primitives: Building institutional infrastructure for humanities e-science. *Future Generation Computer Systems*, 29(2):654–661.

de Jong, F. M. G., Oard, D. W., Heeren, W. F. L., and Ordelman, R. J. F. (2008). Access to recorded interviews: A research agenda. *ACM Journal on Computing and Cultural Heritage (JOCCH)*, 1(1):3:1–3:27.

Dörk, M., Carpendale, S., and Williamson, C. (2011). The information flaneur: A fresh look at information seeking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1215–1224, New York, NY, USA. Association for Computing Machinery.

Garofolo, J. S., Auzanne, C. G. P., and Voorhees, E. M. (2000). The trec spoken document retrieval track: A success story. In *Content-Based Multimedia Information Access - Volume 1*, RIAO '00, pages 1–20, Paris, France, France. Le Centre De Hautes Etudes Internationals D'Informatique Documentaire.

Goldman, J., Renals, S., Bird, S., De Jong, F., Federico, M., Fleischhauer, C., Kornbluh, M., Lamel, L., Oard, D. W., Stewart, C., et al. (2005). Accessing the spoken word. *International Journal on Digital Libraries*, 5(4):287–298.

Graham, S., Milligan, I., and Weingart, S. (2015). *Exploring Big Historical Data: The Historian's Macroscope*. World Scientific Publishing Company.

Gustman, S., Soergel, D., Oard, D., Byrne, W., Picheny, M., Ramabhadran, B., and Greenberg, D. (2002). Supporting access to large digital oral history archives. In *Proceedings of the second ACM/IEEE-CS joint conference on Digital libraries - JCDL '02*, page 18, New York, New York, USA. ACM Press.

Heeren, W., van der Werff, L., de Jong, F., Ordelman, R., Verschoor, T., van Hessen, A., and Langelaar, M. (2009). Easy listening: Spoken document retrieval in choral. *Interdisciplinary science reviews*, 34(2-3):236–252. 10.1179/174327909X441135.

Jammalamadaka, N., Zisserman, A., Eichner, M., Ferrari, V., and Jawahar, C. V. (2012). Video retrieval by mimicking poses. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, pages 1–8.

Kemman, M., Scagliola, S., de Jong, F., and Ordelman, R. (2013). Talking with scholars: Developing a research environment for oral history collections. In *International Conference on Theory and Practice of Digital Libraries*, pages 197–201. Springer.

Kuhn, V., Simeone, M., Marini, L., Bock, D., Craig, A. B., Diesendruck, L., and Satheesan, S. P. (2014). MOVIE: Large Scale Automated Analysis of MOVing ImagEs. In *Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment*, XSEDE '14, pages 1–3, New York, NY, USA. Association for Computing Machinery.

Lincoln, M., Corrin, J., Davis, E., and Weingart, S. (2020). CAMPI: Computer-Aided Metadata Generation for Photo archives Initiative.

Manovich, L. (2013). Visualizing Vertov. *Russian Journal of Communication*, 5(1):44–55.

Melgar-Estrada, L., Koolen, M., Beelen, K., Huurdeman, H., Wigham, M., Martinez-Ortiz, C., Blom, J., and Ordelman, R. (2019). The clariah media suite: a hybrid approach to system design in the humanities. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pages 373–377.

Olesen, C. G. and Kisjes, I. (2018). From Text Mining to Visual Classification: Rethinking Computational New Cinema History with Jean Desmet's Digitised Business Archive. *TMG Journal for Media History*, 21(2):127–145.

Ordelman, R., Melgar, L., Van Gorp, J., and Noordegraaf, J. (2019). Media suite: Unlocking audiovisual archives for mixed media scholarly research. In *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8-10 October 2018*, number 159, pages 133–143. Linköping University Electronic Press.

Ordelman, R. J. and de Jong, F. M. (2011). Distributed access to oral history collections: Fitting access technology to the needs of collection owners and researchers. In *Digital Humanities 2011: Conference Abstracts*. Stanford University Library.

Oyallon-Kololski, J. and Williams, M. (2018). Annotating FloLo: Utilizing Laban Movement Analysis in The Media Ecology Project. In *Women and Silent Screen*, pages 64–70, Shanghai. China Film Press.

Pearlman, K. (2009). *Cutting Rhythms Shaping the Film Edit*. Focal Press, Burlington, MA.

Ruszev, S. (2018). Rhythmic Trajectories – Visualizing Cinematic Rhythm in Film Sequences. *Apparatus. Film, Media and Digital Cultures of Central and Eastern Europe*, 0(7).

Saleh, B., Abe, K., Arora, R. S., and Elgammal, A. (2016). Toward automated discovery of artistic influence. *Multimedia Tools and Applications*, 75(7):3565–3591.

Segel, E. and Heer, J. (2010). Narrative visualization: Telling stories with data. *IEEE transactions on visualization and computer graphics*, 16(6):1139–1148.

Smeulders, A., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380.

Szeto, K.-Y. (2014). *Theater and Film*. Oxford University Press.

Unsworth, J. (2000). Scholarly primitives: What methods do humanities researchers have in common, and how might our tools reflect this. In *Symposium on Humanities Computing: Formal Methods, Experimental Practice. King's College, London*, volume 13, pages 5–00.

Wevers, M. and Smits, T. (2020). The visual digital turn: Using neural networks to study historical images. *Digital Scholarship in the Humanities*, 35(1):194–207.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., and Girshick, R. (2019). Detectron2. https://github.com/facebookresearch/detectron2.