Computational Tools and Methods for Film and Video Analysis

Manuel Burghardt[a], John Bateman[b], Eric Müller-Budack[c] and Ralph Ewerth[d]

[a] https://orcid.org/0000-0003-1354-9089, [b] https://orcid.org/0000-0002-7209-9295,
[c] https://orcid.org/0000-0002-6802-1241, [c] https://orcid.org/0000-0003-0918-6297

Abstract In this chapter, we provide an overview of computational tools and methods for film and video analysis. After introducing the evolution of methods in this field, we go on to lay some theoretical foundations for empirical video analyses. As we focus on using state of the art deep learning methods, we also provide an overview of the types of information that can already be extracted with these methods. Furthermore, we introduce an easy-to-use tool for video analytics, called TIB AV-Analytics (TIB AV-A). We demonstrate how TIB AV-A can be utilised to support the exploration of narrative patterns in the popular TV series *Game of Thrones.* Finally, we conclude by summarising the current state of available tools and methods for computational video analysis and outline some challenges that lie ahead.

Keywords computation film analysis, computational video analysis, distant viewing

1. A Short History of Computational Methods in Film and Videos Studies

Traditionally, the field of Digital Humanities (DH) has a strong focus on textual material, with its origins traced back to Roberto Busa's *Index Thomisticus.* However, in recent years, there has been a growing interest within DH towards film and video (Burghardt & Wolff 2016; Sittel 2017; Heftberger 2018; Burghardt et al. 2020; Arnold & Tilton 2022), leading to the establishment of dedicated special interest groups at both national[1] and international[2] levels. Given the highly interdisciplinary nature of DH, multiple views have emerged regarding the study of film. Burghardt et al. (2020) identify three key perspectives: (1) an *infrastructural perspective* encompassing GLAM (Galleries, Libraries, Archives, Museums) institutions and film archives, (2) a *media*

[1] DHd AG Film & Video: https://dig-hum.de/ag-film-und-video (Accessed: 22 June 2024).

[2] ADHO Special Interest Group AudioVisual material in Digital Humanities: https://avindhsig.wordpress.com/ (Accessed: 22 June 2024).

Manuel Burghardt, John Bateman, Eric Müller-Budack and Ralph Ewerth: Computational Tools and Methods for Film and Video Analysis, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 147–173. DOI: https://doi.org/10.11588/heibooks.1521.c21945

perspective addressing digital encounters in film and media studies, and (3) a *computational perspective,* which is focused on *multimedia information retrieval* and *multimodal information extraction.* As an increasing amount of video material is becoming available in digital form, the computational perspective has gained significant relevance, leading to the emergence of terms such as "distant viewing" (Arnold & Tilton 2019), "distant watching" (Howanitz 2015) and "deep watching" (Bermeitinger et al. 2019), to describe these developments. Wevers & Smits (2020) even propose a "visual digital turn", driven by the capabilities of deep learning techniques.

While the conceptualisation of distant viewing and the like has gained recent attention, early examples of quantitative film studies with a focus on shot analyses can be found with Barry Salt (1974; 2006) and Yuri Tsivian's *Cinemetrics* database (2009). Additionally, other quantitative approaches to stylistic and formal feature analysis have been explored. These include, for instance, language analysis (Hoyt et al. 2014; Byszuk 2020; Bednarek 2023) and colour analysis (Burghardt et al. 2018; Pause & Walkowski 2018; Flueckiger & Halter 2020). Another branch of quantitative analysis of large video corpora has its focus on information visualisation. Manovich's (2013) *Visualizing Vertov* project represents an early example that heavily relies on visualisations to uncover patterns in extensive collections of images and videos. In addition to such visual analytics approaches, a wide range of tools is available for the annotation and analysis of videos and movies. Examples of these tools include *ELAN* (Wittenburg et al. 2006), *Videana* (Ewerth et al. 2009), *ANVIL* (Kipp 2014), and *VIAN* (Halter et al. 2019). For a comprehensive overview of these tools, along with their specific features and functionalities, we recommend the survey paper by Pustu-Iren et al. (2020). Furthermore, there are more recent tools that go beyond the analysis of visual aspects alone and encompass language (spoken and written) as well as audio (music and sound). Notable examples of such tools include Zoetrope (Tseng et al. 2023; Liebl & Burghardt 2023) and TIB AV-A (Springstein et al. 2023).

With this chapter we aim to provide an introduction to computer-assisted, empirical analysis of film and videos. In Section 2, we lay some theoretical foundations for such empirical analyses. Section 3 is about providing an overview of the types of information that can already be extracted using deep learning methods. As setting up such methods can be challenging without advanced technical knowledge, we also introduce an easy-to-use tool called TIB AV-Analytics (TIB AV-A, Section 4). Furthermore, we present a case study in Section 5, demonstrating how TIB AV-A can be utilised to support the exploration of narrative patterns in the popular TV series *Game of Thrones.* In Section 6, we conclude by summarising the current state of available tools and methods for computational video analysis and outline some challenges that lie ahead.

2. Theoretical foundations of empirical film analysis

The study of film has long employed a combination of qualitative and quantitative methodologies (Korte 2004). While qualitative approaches abound in the analysis of movies, as demonstrated by works such as Stam (2000), Prince (2007), Sikov (2010), Ryan & Lenos (2020), and many others, DH approaches to film studies tend to emphasise quantitative and empirical aspects. In this section, we aim to establish some theoretical foundations that contextualise empirical film studies further and relate such endeavours to broader analytic concerns.

A host of *external* engagements with film continue to be broadly relevant for, and may interact with, approaches to film within DH. These range from archiving, historical studies, production studies, investigations of the effects of developing display technologies, provision for information-retrieval from audiovisual data, reception studies (from psychological studies to reviews and critique), to cultural studies of cinema as an institution. In all cases, obtaining a tighter analytic hold *internally,* i.e., of film as artefacts with specific designs for aesthetic or other purposes, can be seen as an important step in understanding the medium. Such an analytic hold is arguably best achieved with the support of empirical studies whereby properties of films are successively revealed and used to draw ever deeper generalisations concerning their functioning. DH approaches to film thus typically seek research strategies that involve both quantitative studies of the distributions and patterning of measurable filmic properties and qualitative, more hermeneutic interpretations of those distributions and patterns (Flückiger 2011; Heftberger 2018).

The primary challenge to be faced in this context is then how to gain access to those details of highly complex audiovisual filmic artefacts that are relevant for interpretation and analysis. Since many questions raised concerning film are interpretative and hermeneutic in nature, it is by no means self-evident how quantitative approaches may support such concerns. This is, indeed, a very general philosophical point raised for many branches of the DH. In the case of film, early quantitative approaches drew much of their motivation from stylometry for literature, but were limited to manual methods restricting both the scale of studies and the kinds of features that could be considered. The most extensive programme of this kind, pursued for many years by Barry Salt (1974; 2007), consequently counted shot lengths and shot scales for selected portions of collections of films drawn from distinct periods and directors. Work in this tradition continues and has revealed a host of historical and regional developments; Cutting & Candan (2015), for example, report on an again predominantly manual analysis of shot lengths of 9,400 English-language films and 1,550 non-English language films released between 1912 and 2013. Similar broad historical changes have been reported for brightness, colour, and shot transitions (e.g., Cutting et al. 2011a; Cutting et al. 2011b). Redfern (2022b) sets out a host of practically applicable R-scripts (R Core Team 2016) that support these kinds of studies.

Heftberger (2018) pursues a different approach by relying on direct visualisations composed of still images taken from films, particularly focusing on the works of Dziga Vertov. Direct visualisations of perceptible features, such as brightness, colour, and so on, have been explored by several researchers in DH since the generation of such visual representations is relatively straightforward. It is, however, questionable to what extent relying simply on (typically) visual perception is an effective method for revealing deeper patterns of interest. In many respects, this is symptomatic of the current highly exploratory phase of DH film studies, where what is used for analysis is what is technologically feasible. To date, more extensive approaches that address a broader range of film phenomena remain limited. Bakels et al. (2020), for example, report on highly detailed, multitracked analyses of films that specifically target issues concerning the audiovisual construction of affect in films drawing on the AdA ontology of film analytic terms[3]. However, these are, again, still largely manual although automated digital techniques are now being successively added as well.

It is evident that empirical distributional studies benefit considerably from technological support so that time-consuming and error-prone manual analysis can be reduced to a minimum, but basic issues remain concerning the overall utility of such approaches to film. For further developments it will be important to relate the capabilities of computationally-supported analysis tools more closely to the research questions raised in the humanistic study of film. To date, there is still a gap to be bridged here. In Heftberger's case, motivation is provided by the fact that, as she notes, Vertov himself paid close attention to formal design features when constructing his films and so the formal analysis is well justified. It is less clear, however, that this can be adopted as a general guideline concerning how to go about film analysis empirically using computational and other quantitative methods.

One of the ways in which more general orientations are currently being developed draws much from a multimedia *information-retrieval* view of computational film analysis. Kurzhals et al. (2016), for example, seek to raise *visual movie analytics* to a more semantic level by combining a variety of sources of information (including scripts and subtitles) to answer the four basic questions of summarisation: *who, what, where,* and *when.* A rich range of automatic processing techniques are combined in this architecture to deliver descriptions of scenes and their events, thereby picking up from a quite different perspective some of the most traditional approaches to film analysis in terms of detailed shot-by-shot descriptions known as film or sequence protocols (cf. Kanzog 1991, 136–151; 163–183). Producing such protocols is extremely time consuming, however, and so any contribution to their automatic construction constitutes a useful advance; we will see many capabilities of these kinds emerging in the years to come and several significant developments in this direction are discussed below. Nevertheless, and more specifically for film as aesthetic artefacts, it is striking that the four basic summarisation questions omit one question that is crucial for film

[3] See https://projectada.github.io/ontology (Accessed: 22 June 2024).

analysis: that is the *how* question. It is often insufficient to describe simply the bare plot structure of a narrative. For film we are often equally interested in the manner of the story telling since this is what drives film as a culturally effective communication form. The inclusion of such information was one of the primary reasons why traditional film protocols were so labour-intensive to produce.

This then returns us directly to several fundamental issues concerning the more interpretative nature of film analysis mentioned above and which DH still needs to come more to grips with. Since it is in general by no means straightforward to relate formal features of film with interpretations, principled ways of raising levels of abstraction to bridge the gap are needed. Vonderau (2017), for example, sets out a useful summary of some of the engagements between more digitally-oriented approaches to film as data on the one hand, and the traditional concerns of film studies on the other. But these questions raised anew within DH also have a longer history as a reoccurring point of criticism made by film scholars with respect to purely quantitative approaches in general. As David Bordwell and colleagues noted very early on, even though it is possible to compute "general statistical norms" in the style of Barry Salt, "such abstractions mean little without a concept of the range of paradigmatic choice dominant at a given period" (Bordwell et al. 1985, 60). Moreover, it is necessary to come to terms with an essential indirectness in the process of meaning attributions in the medium of film that is readily missed in quantitative approaches:

> Sometimes we're tempted to assign absolute meanings to angles, distances, and other qualities of framing. …The analysis of film as art would be a lot easier if technical qualities automatically possessed such hard-and-fast meanings, but individual films would thereby lose much of their uniqueness and richness. The fact is that framings have no absolute or general meanings. (Bordwell & Thompson 2008, 192)

It is only within particular systems of contrasts that specific meanings can be assigned at all (Branigan 1984, 29). Support for investigating this aspect of filmic meaning-making is only provided in current computational tools via interactive interfaces, since the computational processes themselves are not yet in a position to provide reliable hypotheses concerning interpretation. It is then, as we will see in subsequent sections, sensible to provide ready access to automatically captured formal features of film, but still leave the task of assigning interpretations to combinations of these features to human analysts by formulating appropriate queries by hand over combinations of automatically tagged categories (cf., e.g., Kurzhals et al. 2016).

Supporting interpretative tasks more directly will no doubt be a major area of development in the future. For this it will be useful to construct more robust theoretical frameworks that rely on explicit semiotic foundations more than has been the case in DH previously. Such foundations will need to align with contemporary views of semiotics that, on the one hand, are equally supportive of quantitative and qualitative kinds of description and, on the other, are capable of spanning both the recognition of more formal technical features and the entire process of hermeneutic interpretation in context. Only then can studies begin to balance the purely bottom-up, data-driven approaches that currently predominate (Redfern 2022b).

One such account of semiotics that has been related both to DH and to the analysis of film is set out in Bateman et al. (2017). In this approach expressive resources of a medium, which for film include montage, lighting, colour, music, and many more, are each characterised at three distinct levels of abstraction: *material* (supporting measurements), *form* (supporting explicit representations of paradigmatic systems of contrasts), and *discourse* (supporting interpretation in context). It is the latter level of description that moves the account beyond the capabilities of the more structural semiotics that became prevalent in the 1960s. The distinct levels of abstraction provided are directly relevant for DH because they motivate distinct classes of annotations that may be deployed to describe any artefacts being analysed and so serve to organise larger scale collections (Bateman 2022).

The multi-levelled view is now also essential for incorporating the new generation of computational techniques based on deep learning within a coherent overall framework. Such techniques are no longer limited to operating within single levels of abstraction and deliver useful results ranging from low-level formal features of film through to direct descriptions of semantic content. Semiotically, therefore, these components function similarly to the linguistic notion of constructions, which typically combine information from diverse levels of abstraction to offer reusable building-blocks for communication (Goldberg 1995). Specifically filmic constructions, often termed filmic idioms, are now also receiving formal treatments (Wu et al. 2018). The logical next step is therefore to combine these descriptions with automatic analysis components for prediction and recognition, and to support both visualisation and statistical evaluations building on the higher levels of abstraction achieved. Preliminary steps in this direction will be suggested below.

3. A short introduction to multimodal information extraction frameworks

Videos comprise multiple expressive modalities, such as image, audio (including speech), and text (overlaid text in video frames). To analyse individual videos and entire corpora, information from all modalities is required. Since the manual analysis of films is a very time-consuming task, there is a large need for automatic pattern recognition and multimedia retrieval methods to support DH researchers. In recent years, tremendous progress has been achieved in computer science fields such as computer vision, audio analysis, and natural language processing, thanks to deep learning models and the availability of large-scale datasets for training. In this section, we provide an overview of approaches for information extraction from images, audio, and text for film analysis, although it must be noted as well that this is only a selection of approaches made from those that we considered most relevant for video analysis, driven by our joint collaborations with researchers from the DH (i.e., film analysts, semioticians, media and communication scientists). In addition, there are further multimodal approaches that combine the capabilities of the methods listed here.

Computer vision: Different aspects of visual information are important for film analysis, ranging from low-level features (e.g., colour, brightness), over camera settings (e.g., shot scale, camera movement), to more complex information (e.g., actions, places, persons). There are many libraries (e.g., *scikit-learn*[4]) to extract low-level features such as *brightness, colour,* and *contrast.* For most other computer vision tasks, deep learning models such as convolutional neural networks (e.g., He et al. 2016) or transformers (e.g., Radford et al. 2021) are typically used. Methods for *temporal video segmentation* are an essential step to structure a video and can be categorised with regard to *shot* (e.g., Souček & Lokoč 2020), *scene, story, and topic boundary detection* (e.g., Wu et al. 2023). Another relevant information is the *camera setting* (e.g., shot scale, camera movement), camera pose (i.e., pitch, yaw, roll), and camera angle (e.g., Liu et al. 2022) that can be predicted by deep learning models. Approaches for *optical character recognition* (e.g., Kuang et al. 2021) automatically detect overlaid text in images that can be further analysed with approaches for natural language processing (see below). There are various deep learning approaches for image content analysis. In particular, the identification of *persons* (e.g., Deng et al. 2020), *facial attributes* (e.g., emotions, head pose, gender; Hempel et al. 2022; Serengil & Ozpinar 2021), and other *concepts* (e.g., animals, cars, objects; Radford et al. 2021) has been well-studied by the computer vision community. Furthermore, approaches for the identification of *place categories* (e.g., church, market, restaurant; Zhou et al., 2018), *geographical locations* (e.g., Müller-Budack et al. 2018; Theiner et al. 2022), and *events* (e.g., protests, elections, natural disasters; Müller-Budack et al. 2021) have been presented that can be used to categorise and characterise film segments. While deep learning models are often explicitly optimised for such tasks using labelled training data, recent vision-language models such as *CLIP* (*Contrastive Language-Image Pretraining;* Radford et al. 2021) have been trained with hundreds of millions of image-text pairs to implicitly learn visual concepts. These models can be applied to many tasks since they can measure the *similarity of arbitrary concepts* (e.g., objects, weather, occupation) to an image based on a textual description (i.e. a prompt). Recently, novel large vision-language models (e.g., Alayrac et al. 2022; Dai et al. 2023) combine the capabilities of these approaches with large language models such as OpenAI's GPT-4[5] (*Generative Pre-training Transformer* 4) and achieve impressive results for many applications,

[4] See https://scikit-learn.org (Accessed: 22 June 2024).

[5] See https://openai.com/gpt-4 (Accessed: 22 June 2024).

including film and video analysis (Zhang et al. 2023). While most aforementioned approaches focus on single images and have to be applied to each video frame, methods for video classification also consider temporal context from frame sequences (e.g., Ni et al. 2022) for further applications such as *action recognition* (e.g., running, talking).

Audio analysis: Basic analysis steps for audio regard *low-level features,* such as *amplitude, volume,* and *spectrogram* (e.g., using the *librosa* library for *Python*) that can indicate volume changes, (rhythmic) patterns, music, and other sound effects. The *transcription of the spoken language* is another highly relevant task for film analysis. Recently, neural transformer architectures have been introduced for automatic speech recognition (e.g., *Whisper;* Radford et al. 2023) that achieve impressive results across many languages.[6] Automatically extracted transcripts enable an in-depth analysis of speech using tools from natural language processing (see below). Methods for *speaker diarisation* (e.g. Bredin & Laurent, 2021) can further refine the speech transcript by assigning the identity of the corresponding speaker to, for example, analyse the spoken language for each speaker individually or to find forms of conversations (e.g., monologue, dialogue) in a film. It also serves as the basis for the *identification of voice characteristics,* such as gender (e.g., Baevski et al. 2020) and emotions (e.g., Ravanelli et al. 2021). Besides the analysis of speech, researchers have also focused on the *detection* and *classification of music* (e.g., Liu et al. 2021) as well as more general *audio classification* (e.g., Wu et al. 2022). Motivated by CLIP (see above), CLAP (*Contrastive Language-Audio Pretraining;* Wu et al. 2022) has been trained with several hundred-thousand audio and text pairs to enable classification of *arbitrary* audio concepts (e.g., sound events like *siren wailing* or *rain falling*) based on textual prompts.[7]

Natural language processing: As mentioned above, methods for optical character recognition from images (video frames) and automatic speech recognition from audio allow for the extraction of textual information from videos based on overlaid text and speech. Methods from natural language processing enable many perspectives to further work with such language data. For example, *part-of-speech tagging* (e.g., *spaCy*[8]) can be applied for syntax analysis, to better understand the grammatical structure of a sentence. *Named entity recognition* and *disambiguation* (e.g., *spaCy,* Wu et al. 2020) can automatically detect mentions of persons, locations, and events that play crucial roles in videos and films.[9] Moreover, there are numerous approaches for the classification of *topics* (e.g., Grootendorst 2022) and *sentiment* (e.g., Devlin et al. 2019) that can provide insights to the overall plot as well as the emotional tone and

[6] See https://github.com/openai/whisper#available-models-and-languages (Accessed: 22 June 2024).

[7] Cf. the chapter from Ch. Weiß in this volume.

[8] See https://spacy.io (Accessed: 22 June 2024).

[9] Cf. the chapter from E. Gius in this volume.

dynamics of characters throughout the film.[10] Very recently, large language models such as OpenAI's GPT-4 have been massively applied to a wide range of the aforementioned tasks and beyond.

4. Video analysis with the TIB AV-Analytics (TIB AV-A) Tool

The implementation of the deep learning techniques introduced in the previous section can pose substantial technical challenges. As an intermediate step, a number of toolkits[11] have been proposed that provide a basic layer of abstraction, but still require advanced technical knowledge and data literacy. However, to make available the advantages of large-scale pattern recognition and multimedia retrieval methods (see Section 3) to the broader community of scholars that work with audiovisual material, an easy-to-use tool with a graphical user interface is desirable. This is the main motivation for the TIB AV-Analytics platform (TIB AV-A[12]) that is currently being developed by TIB – Leibniz Information Centre for Science and Technology, in collaboration with film scholars from the University of Mainz.

TIB AV-A is a web-based platform for systematic film and video analysis (a screenshot is shown in Fig. 1). The platform uses modern web technologies and a plugin structure to simplify the integration of new plugins for developers and researchers to maintain TIB AV-A at the current state of the art. We use containers (e.g., *Docker*[13]) for virtualization for easy setup and to manage software dependencies, as well as an inference server (currently *Ray*[14]) for stable deployment. To ensure interoperability with other video analysis tools, TIB AV-A provides an Application Programming Interface (API) and import and export of results in common data formats, such as csv *(comma separated values)* files, as well as to the widely used ELAN video annotation tool (Wittenburg et al. 2006). The source code is publicly available.[15] More details are described by Springstein (2023).

In contrast to previous video analysis tools that either only allow for manual annotations (e.g., *ANVIL*[16], by Kipp 2014; *Cinemetrics*[17], by Tsivian 2009; *ELAN*[18], by

[10] Cf. the chapters from M. Althage and R. Sprugnoli in this volume.

[11] *Distant viewing toolkit,* Python notebooks (Arnold & Tilton 2020): https://github.com/distant-viewing/dvt; *Computational Film Analysis with R* (Redfern 2022b): https://cfa-with-r.netlify.app/index.html (Both accessed: 22 June 2024).

[12] See https://service.tib.eu/tibava (Accessed: 22 June 2024).

[13] See https://www.docker.com (Accessed: 22 June 2024).

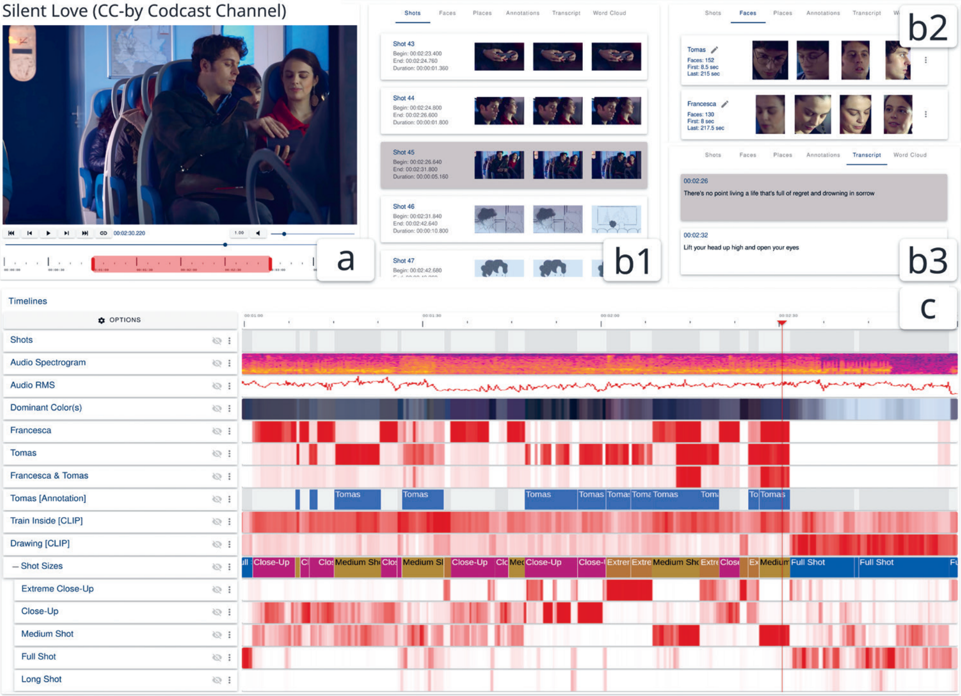
[14] See https://www.ray.io (Accessed: 22 June 2024).

[15] See https://github.com/TIBHannover/tibava (Accessed: 22 June 2024).

[16] See http://www.anvil-software.de (Accessed: 22 June 2024).

[17] See https://cinemetrics.uchicago.edu (Accessed: 22 June 2024).

[18] See https://archive.mpi.nl/tla/elan (Accessed: 22 June 2024).



Shot 43
Begin: 00:02:23.400
End: 00:02:24.760
Duration: 00:00:01.360

Shot 44
Begin: 00:02:24.800
End: 00:02:26.600
Duration: 00:00:01.800

Shot 45
Begin: 00:02:26.640
End: 00:02:31.800
Duration: 00:00:05.160

Shot 46
Begin: 00:02:31.840
End: 00:02:42.640
Duration: 00:00:10.800

Shot 47
Begin: 00:02:42.680

Tomas
Faces: 152
First: 8.5 sec
Last: 215 sec

Francesca
Faces: 130
First: 8 sec
Last: 217.5 sec

00:02:26
There's no point living a life that's full of regret and drowning in sorrow

00:02:32
Lift your head up high and open your eyes

Fig. 1 Interface of TIB AV-A for the short movie *Silent Love* (CC-by Codcast Channel, Original Video: https://www.youtube.com/watch?v=KuuEs0oVVS8 [Accessed: 22 June 2024]). It contains a video player (a), an overview of detected shots (b1), persons (b2), and the speech transcript (b3). The timelines (c) can display categorical (e.g., "*Tomas*") and numerical values (e.g., "*Drawing [CLIP]*"). Timelines with numerical values indicate, e.g., the probability whether a concept is depicted in a video. The user can select the visualisation type (line chart, colour chart) and colour (here: from white [unlikely] to red [likely]).

Wittenburg et al. 2006) or contain only a few selected methods for automatic content analysis (e.g., *Videana*, by Ewerth et al. 2009; *VIAN*[19], by Halter et al. 2019), TIB AV-A provides a vast collection of state-of-the-art pattern recognition approaches without the necessity of advanced technical knowledge or specific hardware requirements. Users from various disciplines can simply upload their own videos and then have access to a variety of analytical perspectives. An overview of currently supported methods for filmic analysis is provided in Tab. 1.

[19] See https://www.vian.app (Accessed: 22 June 2024).

Tab. 1 Overview of current methods for image and video analysis as well as audio and speech analysis in TIB AV-A.

Image and video analysis	Basic image features: dominant **colour(s) and brightness**
	Shot boundary detection
	Cut frequency (cf. Redfern 2022a), i.e., the frequency of shot transitions
	Shot scale classification, to differentiate between the following shot scales: extreme close-ups, close-ups, medium shots, full shots, and long shots
	Place classification (e.g., church, market, restaurant, etc.)
	Person recognition based on an example image
	Place and person clustering to automatically find the most frequently appearing places and persons/actors
	Facial expression recognition (e.g., angry, happy)
	Zero-shot image classification for arbitrary visual concepts based on textual descriptions (e.g., "A photo taken in a train", see Fig. 1)
	Zero-shot video classification for arbitrary audiovisual concepts based on textual descriptions (e.g., "A video with celebrating people")
	Image captioning to automatically describe frames within a video
Audio and speech analysis	Basic audio features: **amplitude curve** (waveform), **volume** (root mean square), and the **frequency spectrum**
	Speech recognition to automatically transcribe speech in videos

Besides some standard analysis tasks (e.g., colour analysis, shot boundary detection), the addition of speech recognition and zero-shot image and video classification is most notable in TIB AV-A. High-quality transcripts (i.e., with a low word error rate) enable a much better analysis of speech using approaches from natural language processing for tasks like *topic modeling*, *named entity linking*, etc., which will be added in TIB AV-A in the future. Furthermore, zero-shot image and video classification enable various downstream tasks. Based on a textual prompt, the underlying vision language models, i.e., *CLIP* (Radford et al. 2021) and *InstructBLIP* (Dai et al. 2023), can recognise (a set of) *arbitrary* concepts. In this way, users can automatically search videos for various concepts ranging from real-world objects (e.g., flags, cars, etc.) and animals over environmental settings (e.g., places, weather, daytimes) to much more complex concepts, e.g., occupations of persons (e.g., police officer, reporter), events (e.g., natural disasters, demonstrations, types of sports), etc.

Although TIB AV-A provides a large set of state-of-the-art methods for automatic film analysis, DH researchers are often interested in more advanced patterns that can comprise a combination of features. For example, sequences in movies with high *shot density*, sudden *volume changes*, and *close-up shots* may indicate suspenseful key scenes or actions in movies. The combination of features can also add conditions to certain patterns to, for example, search for actions if a specific person or object is visible (see Fig. 1). To enable such combinations, TIB AV-A offers the option to aggregate probabilities of certain features (e.g., scenes, emotions, shot scales) with logical operations *(or/and).* Based on the features extracted from a given video, users can create interactive visualisations for qualitative analysis. Currently, TIB AV-A supports a word cloud visualisation based on the extracted speech transcripts as well as scatter and line plots for which the user can display (and hide) specific features and feature combinations (see Fig. 1). Moreover, graphs that can, for example, show character constellations and their occurences at specific locations and places can be created.

5. Case study: Analysing the end of the *Game of Thrones* series for narrative patterns

We have seen in the previous section how state of the art tools for managing the automatic analysis of films provide support for a variety of analysis methods. The components being integrated into TIB AV-A cover two main kinds: first, the automatic analysis of films with respect to categories and properties that hold for all films, such as shot boundaries, colour ranges, sound spectrograms, and the like, and second, the automatic analysis of films with respect to categories, semantic constructs, or formal features that are selected by the human analyst. In both areas, we can expect the accuracy, precision, and diversity of results delivered to grow substantially in the coming years. Several questions remain, however, concerning how these capabilities can be leveraged to support the distinct kinds of analyses that may be targeted for film. In this section, we show an example of analysis that focuses specifically on uncovering larger-scale filmic structures that serve functional aims such as storytelling.

To make the discussion concrete, analysis will proceed with respect to the closing scenes of the last episode in the final series of *Game of Thrones,* created by David Benioff and D.B. Weiss for HBO and first aired in 2019.[20] The *what* of this segment of material is quickly described: the three main characters from the story's central Stark family, Jon Snow, Arya Stark, and Sansa Stark, begin new stages of their lives. Jon Snow passes the boundary separating civilisation from the icy north, Arya Stark sails west to look for new lands, and Sansa Stark is crowned queen. Thus the series ends. Filmically, however, the presentation of these events deploys a collection of well-known techniques yielding a tightly structured comparison of the respective fates of the individuals depicted. It becomes more relevant from the perspective of

[20] The analysed scene can be seen at: https://www.youtube.com/watch?v=zUZvYAjaEZk (Accessed: 22 June 2024).

film analysis, therefore, to address the specifics of the *how* question concerning the segment's construction.

5.1 Workflow

We will show now how use of the TIB AV-A platform can support exploration of filmic organisation of this kind, revealing first the internal structure of the segment and then discussing briefly how this may be drawn into larger scale investigations of film form. We will also emphasise how working from the aesthetics and poetics of film analysis helps to set implementation priorities for the kinds of features that would be optimally beneficial for progressively moving ever more of the manual and semi-automatic analysis to automatic analysis. In the following workflow we also use the ELAN tool for manual annotation and correction of the automatic annotations and some custom R scripts for plotting of the results.

The first step is to load the film segment of interest into TIB AV-A and to perform the standard automatic processing pipelines of shot segmentation, shot scale, and so on. At this point, items that are known to be of particular relevance for the segment can also be used for specific categories – for example, searching for faces of the principal characters on the basis of uploaded images of their faces or by using natural language phrases for zero-shot content-based segmentation.

The second step is to export the analysis tracks from TIB AV-A and translate them into a form appropriate for further segmentation and manual annotation with ELAN or similar tools; this latter step is performed here locally using specific processing scripts. This allows errors in automatic processing to be corrected and further filmic features to be added that are not yet provided automatically by TIB AV-A. Relevant examples of these in the present case are camera movements since the segment relies extensively on camera movement cohesion across subsequences. The general scheme of analysis then follows that set out in Bateman & Schmidt (2012), where shots are allocated to spatiotemporal regions. Human visual perception is generally very fast and accurate in deciding whether it has encountered a particular place before and this kind of continuity is well-known from psychological studies as a fundamental unit for extended discourse comprehension (Zacks 2010; Loschky et al. 2020). Annotation tracks, or tiers, are consequently defined in ELAN so that shots can be assigned to them as revealed by any of the levels of analysis available. This is an area where increasingly accurate scene recognition combined with visual similarity measures can be expected to provide substantial improvements in the near future for supporting automatic or semi-automatic analysis. By these means we can see how questions driven directly by the needs of film poetic and aesthetic analysis may be progressively taken over and supported by the developing computational tools; required features that are missing may first be added manually and then supported computationally as they become available.

The third, and for current purposes final, step is to export the ELAN analysis further for focused examination of reoccurring filmic patterns. For this we use custom-built R scripts running locally that directly transform ELAN annotations into visualisations of the filmic structure, overlaid with results of automatic and manual analysis as desired. Whereas many classical formal editing features can now be explored in rather sophisticated ways for their statistical properties in R (cf. Redfern 2022b), here we will be concerned more with deriving higher-level organisational properties that often correspond more directly with interpretations. The visualisations employed here are defined in Bateman & Schmidt (2012) and draw loosely on musical notation, laying out successive shots horizontally so that further structural relationships, properties and groupings can be added freely. In short, we attempt here to identify functionally relevant sequences of combinations of filmic features that can start moving us beyond overall statistics of transitions, co-occurrences, and the like (Bateman 2014).

5.2 Analysis

The basic structure of the example segment is then as given in Fig. 2. This shows the shots of the scene running horizontally numbered along the bottom row, together with brief functional descriptions of those shots included for ease of reference along the top row. Whenever what has been identified as a shot responds to distinct functional groupings, it is further divided into 'subshots' – as seen, for example, in shot 15, which further divides into a segment tracking a character walking (15.1) followed by a stationary focus on that character (15.2).

This visualisation readily allows us to see the essentially three-line development of the sequence, where successive shots frequently range across the distinct locales of the three main characters (arrayed vertically). This structure is defined formally in Bateman & Schmidt (2012, 222–226) as a tri-partitioned polyspatial alternation and commonly expresses contrast and comparison. Each shot is also labelled here with its shot scale, running from tight close-up or detail shots (TS) to extra long shots (ELS). The sequence thus starts with three tight shots running successively across Jon Snow's (JS), Arya Stark's (AS), and Sansa Stark's (SS) locales; the next three shots, also tight shots, repeat these transitions in reversed order; and so on.

Filmically it is then interesting to examine more closely how the construction of the segment maintains coherence despite this rapid cross-cutting between scenes. To explore this, we successively augment the visualisation with further layers of information from the annotation. Fig. 3, for example, shows the visualisation with the annotation tracks of distinct kinds of camera use folded in, both as labels and as coloured groupings shown over the affected shots. The shots classified according to *camera direction* in this figure then show well how direction maintains cohesion across the multiple locales. Shots 18–22, for example, maintain rightwards tracking,

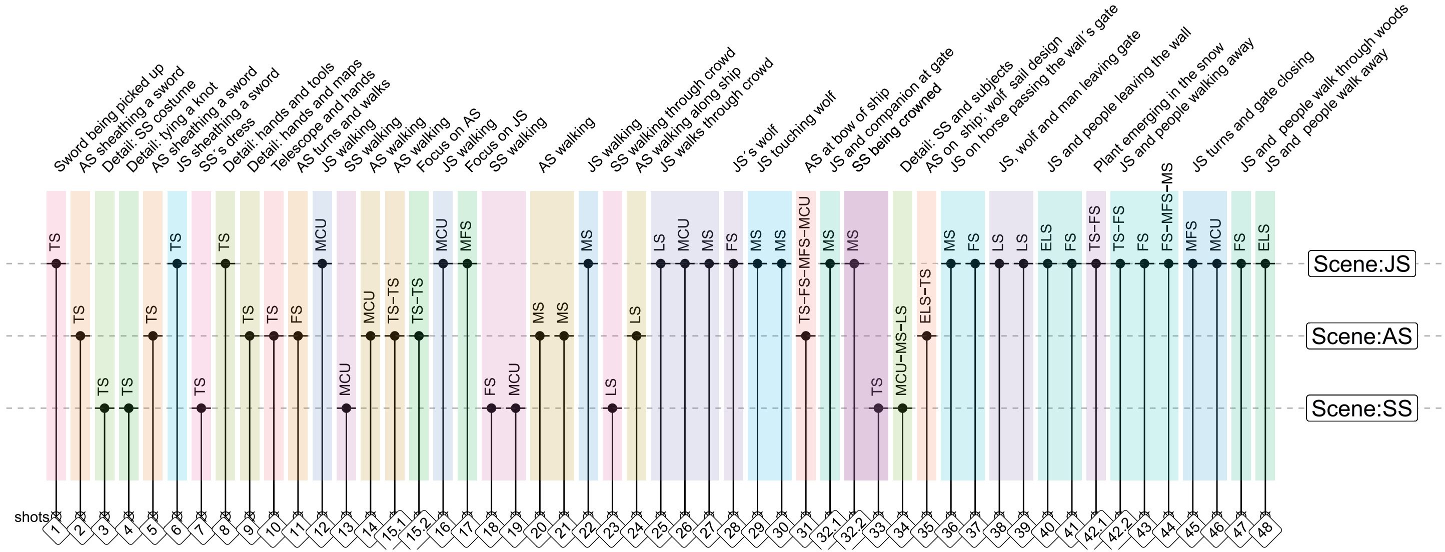


Fig. 2 Basic visualisation of the annotated film structure of the *Game of Thrones* segment; shots run horizontally, numbered at the bottom; colouring indicates semantic content grouping (all graphs produced with the R-package *ggplot*, Wickham 2016). The shot-scale abbreviations are based on standard shot scales, in increasing distance: *tight or detail shot* (TS), *closeup* (CU), *medium closeup* (MCU), *full shot* (FS), *medium full shot* (MFS), *medium shot* (MS), *long shot* (LS), *extra long shot* (ELS).

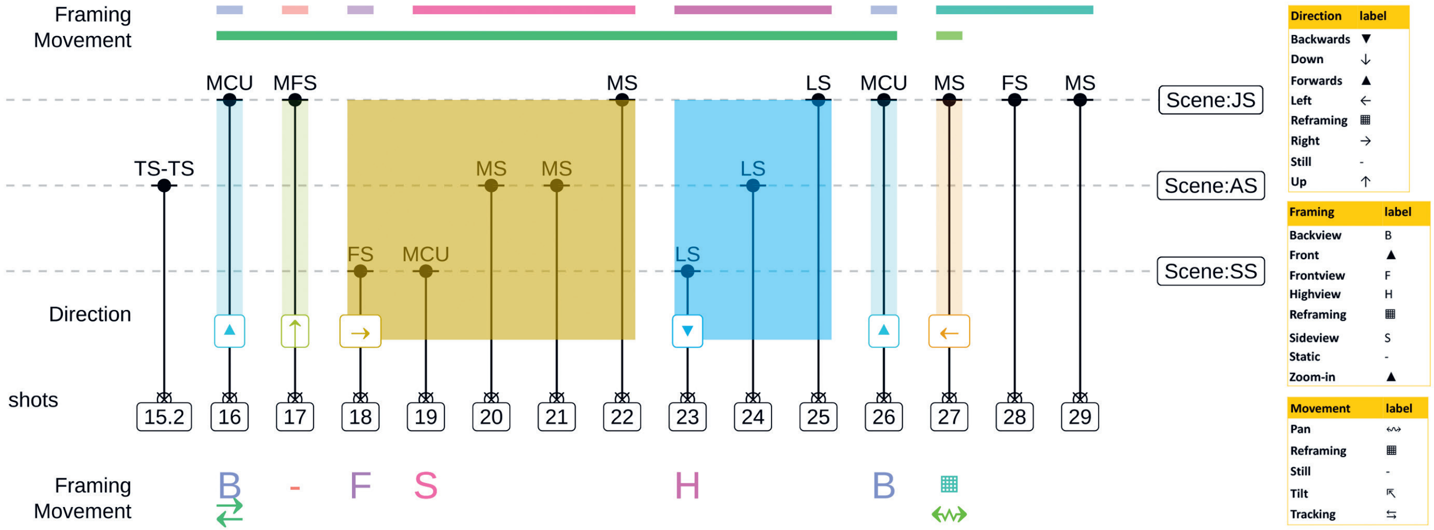


Fig. 3 Visualisation of the annotated film structure of the *Game of Thrones* segment (shots 15 – 29) augmented with camera use information. In this visualisation, camera direction has been prioritised showing grouping with larger coloured blocks. The bars along the top show the grouping that is imposed by framing and movement; the symbols at the bottom show what kind of camera use is involved in each case.

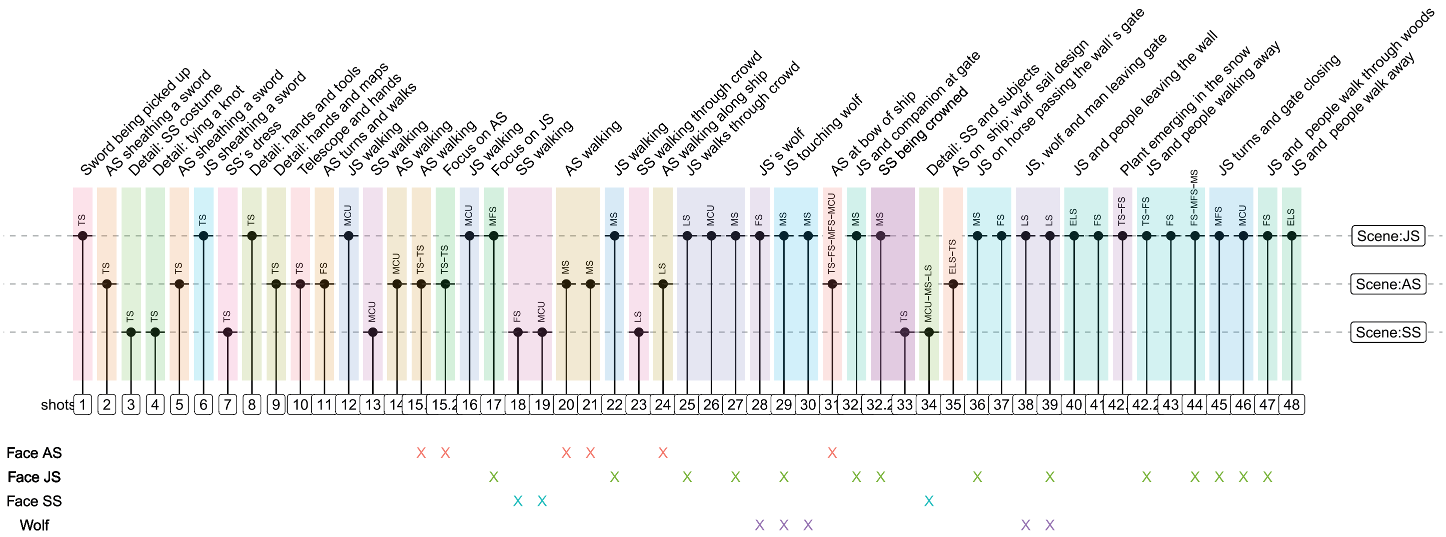


Fig. 4 Visualisation of the annotated film structure of the *Game of Thrones* segment with the results of automatic face recognition and CLIP-based semantic concept recognition.

while shots 23–25 maintain the camera moving out of the scene space. Successive shots within the same locale (shots 28–30, 38–40, 42–46 in Fig. 2), in contrast, appear without any marked camera direction. And, crucially, none of these filmic technical features carries the meaning of the construction alone; it is only in their structural composition that a reliably interpretable form emerges.

The coloured bars running across the top of the graph show similarly how other dimensions of camera use, here framing and movement, also serve to group shots, again often across the three locales. The lower symbol lines of the graph indicate which kinds of framing and movement are playing out more finely; shots 19–22, for example, all exhibit a constant framing (S: "sideview") that contributes to holding the sequence together. Any of these filmic properties can be selected for visual prominence in the graphs so that the differing kinds of structures can be made visible. Relevant examples here would be continuous diegetic sound matching across scenes (such as footsteps) as well as the overall non-diegetic music organising the segment, with the *Stark family motif* running through shots 1–17, gradually blending with the *Game of Thrones* theme across shots 18–30, which then takes over in the remaining shots 31–48.

It is also possible to overlay other automatic annotation results obtained from TIB AV-A. Fig. 4, for example, shows where TIB AV-A recorded one of the three main faces of the characters occurring with high confidence (classified by visual similarity) and occurrences of Jon Snow's wolf (classified by zero-shot semantic classification using CLIP as described above). Here it is interesting for the filmic construction of the sequence that identification of the protagonists is left quite late: only from shot 15 onwards are faces shown as the story moves towards its finale. The placement of the wolf in the middle of the scenes involving Jon Snow is also quite accurate.

While the views shown so far support further exploration of how this sequence is constructed, for the future it will be beneficial to define structural patterns on the basis of the revealed structures that can then in turn be fed back into the automatic search capabilities of TIB AV-A and other tools. This requires the definition of patterns as search queries. For the present case, for example, one would want to search for repeating sequences of shots that are each drawn from a different location but which nevertheless maintain a collection of identical formal technical features, such as camera movement, direction, and so on. Extending such pattern queries to include any of the possible automatically ascertained features promises to dramatically change the state of the art for computer-supported film analysis at scale as well as renewing contact with more hermeneutically-driven research challenges.

6. Conclusion and challenges ahead

We have come a long way in the analysis of film and video in DH. In the era of deep learning advancements, an extensive array of methods has emerged, facilitating the automated extraction of diverse multimodal features. This surge in quantitative data availability necessitates the development of a corresponding analytical framework. We believe that such a framework should be grounded both in empirical standards and in theoretical underpinnings such as multimodal theory and semiotics. It would also benefit from integrating concepts from common taxonomies used by researchers in the DH, such as the AdA film ontology[21], for which first promising experiments have been conducted with TIB AV-A. However, a suitable integration that also captures the hierarchical nature of such ontologies remains to be developed.

While empirical analysis can be seen as a cornerstone of computational film analysis, there's a compelling argument for the integration of exploratory tools like TIB AV-A. So far, most existing tools focus on the exploration and visualisation of single videos. To actually realise the concept of distant viewing (Arnold & Tilton 2019) across multiple videos, we must devise methods for simultaneously visualising multiple videos – an intricate task due to the dynamic nature inherent to video content. First strides have already been taken in this direction through cultural analytics (Manovich 2020) and visual movie analytics (Kurzhals et al. 2016). However, as the trend in textual DH goes towards scalable viewing (Weitin 2017), i.e. a hybrid approach that allows scholars to transition smoothly between close and distant viewing perspectives, this concept holds promise for the analysis of video material as well. Some first examples of scalable viewing can be found in approaches for the visualisation of news videos (Liebl & Burghardt 2023b; Ruth et al. 2023) as well as more generic tools such as *PixPlot*[22] or the *Collection Space Navigator* (Ohm et al. 2023).

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barrerira, R., Vinyalis, O., Zisserman, A., & Simonyan, K. (2022). Flamingo. A Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems 35* (pp. 23716–23736). New Orleans, Louisiana: Neural Information Processing Systems. DOI: https://doi.org/10.48550/arXiv.2204.14198 (Accessed: 22 June 2024).

[21] See https://projectada.github.io/ontology (Accessed: 22 June 2024).

[22] *PixPlot*, from Yale DH Lab: https://github.com/YaleDHLab/pix-plot (Accessed: 22 June 2024).

Arnold, T., & Tilton, L. (2019). Distant viewing. Analyzing large visual corpora, *Digital Scholarship in the Humanities,* 34(1), 3–16. DOI: https://doi.org/10.1093/llc/fqz013 (Accessed: 22 June 2024).

Eid. (2020). Distant Viewing Toolkit. A Python Package for the Analysis of Visual Culture, *Journal of Open Source Software,* 5(45).

Eid. (2022). Analyzing Audio/Visual Data in the Digital Humanities. In J. O'Sullivan (Ed.), *The Bloomsbury Handbook to the Digital Humanities* (pp. 179–187). London: Bloomsbury Publishing.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0. A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems 33* (pp. 12449–12460). Online: Neural Information Processing Systems. DOI: https://doi.org/10.48550/arXiv.2006.11477 (Accessed: 22 June 2024).

Bakels, J.-H., Grotkopp, M., Scherer, T., & Stratil, J. (2020). Digitale Empirie? Computergestützte Filmanalyse im Spannungsfeld von Datenmodellen und Gestalttheorie, *Montage AV – Zeitschrift Für Theorie Und Geschichte Audiovisueller Kommunikation,* 29(1), 99–118.

Bateman, J. A. (2014). Looking for what counts in film analysis. A programme of empirical research. In D. Machin (Ed.), *Visual Communication* (pp. 301–330). Berlin/Boston: De Gruyter Mouton.

Id. (2022). Growing theory for practice. Empirical multimodality beyond the case study, *Multimodal Communication,* 11(1), 63–74. DOI: https://doi.org/10.1515/mc-2021–0006 (Accessed: 22 June 2024).

Id., & Schmidt, K.-H. (2012). *Multimodal Film Analysis. How Films Mean.* London: Routledge.

Bateman, J., Wildfeuer, J., & Hiippala, T. (2017). *Multimodality. Foundations, Research and Analysis. A Problem-Oriented Introduction.* Berlin/Boston: De Gruyter Mouton.

Bednarek, M. (2023). *Language and Characterisation in Television Series. A corpus-informed approach to the construction of social identity in the media.* Amsterdam: John Benjamins Publishing Company [= *Studies in Corpus Linguistics,* 106].

Bermeitinger, B., Gassner, S., Handschuh, S., Howanitz, G., Radisch, E., & Rehbein, M. (2019). Deep Watching. Towards New Methods of Analyzing Visual Media in Cultural Studies. In *Book of Abstracts of the International Digital Humanities Conference (DH).* Utrecht: Alliance of Digital Humanities Organizations. DOI: https://doi.org/10.13140/RG.2.2.12763.72486 (Accessed: 22 June 2024).

Bordwell, D., & Thompson, K. (2008). *Film Art. An Introduction.* New York: McGraw Hill.

Eid., & Staiger, J. (1985). *The Classical Hollywood Cinema. Film, Style and Mode of Production to 1960.* New York: Columbia University Press.

Branigan, E. (1984). *Point of View in the Cinema.* Berlin/Boston: De Gruyter Mouton.

Bredin, H., & Laurent, A. (2021). End-To-End Speaker Segmentation for Overlap-Aware Resegmentation. In *Proceedings of the Interspeech 2021* (pp. 3111–3115). Brno: International Speech Communication Association. DOI: https://doi.org/10.21437/Interspeech.2021-560 (Accessed: 22 June 2024).

Burghardt, M., Heftberger, A., Pause, J., Walkowski, N.-O., & Zeppelzauer, M. (2020). Film and Video Analysis in the Digital Humanities. An Interdisciplinary Dialog, *Digital Humanities Quarterly,* 14(4), 1–37. URL: http://www.digitalhumanities.org/dhq/vol/14/4/000532/000532.html (Accessed: 22 June 2024).

Burghardt, M., Kao, M., & Walkowski, N.-O. (2018). Scalable MovieBarcodes. An Exploratory Interface for the Analysis of Movies. In *Vis4DH. 3rd IEEE VIS Workshop on Visualization for the Digital Humanities.* Berlin: Institute of Electrical and Electronics.

Burghardt, M., & Wolff, Ch. (2016). Digital Humanities in Bewegung. Ansätze für die computergestützte Filmanalyse. In E. Burr (Ed.), *DHd 2016. Modellierung – Vernetzung – Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma. Konferenzabstracts.* 2. überarbeitete und erweiterte Auflage (pp. 191–195). Leipzig: Verband Digital Humanities im deutschsprachigen Raum. URL: https://www.dhd2016.de/sites/default/files/dhd2016/files/boa-2.0_ohne_Vorwort.pdf (Accessed: 22 June 2024).

Byszuk, J. (2020). The Voices of Doctor Who. How Stylometry Can be Useful in Revealing New Information About TV Series, *Digital Humanities Quarterly,* 14(4). URL: http://www.digitalhumanities.org/dhq/vol/14/4/000499/000499.html (Accessed: 22 June 2024).

Cutting, J. E., Brunick, K. L., & DeLong, J. E. (2011a). The changing poetics of the dissolve in Hollywood film, *Empirical Studies of the Arts,* 29(2), 149–169.

Eid., Iricinschi, C., & Candan, A. (2011b). Quicker, faster, darker. Changes in Hollywood film over 75 years, *I-Perception,* 2(6), 569–576.

Cutting, J. E., & Candan, A. (2015). Shot Durations, Shot Classes, and the Increased Pace of Popular Movies, *Projections. The Journal for Movies and Mind,* 9(2), 40–62.

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., & Hoi, S. (2023). InstructBLIP. Towards General-purpose Vision-Language Models with Instruction Tuning [Preprint]. *arXiv.* DOI: https://doi.org/10.48550/arXiv.2305.06500 (Accessed: 22 June 2024).

Deng, J., Guo, J., Liu, T., Gong, M., & Zafeiriou, S. (2020). Sub-center arcface. Boosting face recognition by large-scale noisy web faces. In A. Vedaldi, H. Bischof, T. Brox, & J. M. Frahm (Eds.). *Proceedings of the European Conference on Computer Vision 2020* (pp. 741–757). Cham: Springer [= *Lecture Notes in Computer Science,* 12356]. DOI: https://doi.org/10.1007/978-3-030-58621-8_43 (Accessed: 22 June 2024).

Devlin, J., Chang, M.-W., Kenton, L., & Toutanova, K. (2019). BERT. Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the*

Conference of the North American Chapter of the Association for Computational Linguistics. Human Language Technologies 2019 (pp. 4171–4186). Minneapolis: Association for Computational Linguistics. DOI: https://doi.org/10.18653/V1/N19-1423 (Accessed: 22 June 2024).

Ewerth, R., Mühling, M., Stadelmann, T., Gllavata, J., Grauer, M., & Freisleben, B. (2009). Videana. A Software Toolkit for Scientific Film Studies. In M. Ross, M. Grauer & B. Freisleben (Eds.), *Digital Tools in Media Studies. Analysis and Research. An Overview* (pp. 101–116). Bielefeld: Transcript Verlag.

Flückiger, B. (2011). Die Vermessung ästhetischer Erscheinungen, *Zeitschrift für Medienwissenschaft,* 3(2), 44–60. DOI: https://doi.org/10.1524/zfmw.2011.0022 (Accessed: 22 June 2024).

Ead., & Halter, G. (2020). Methods and Advanced Tools for the Analysis of Film Colors in Digital Humanities, *Digital Humanities Quarterly,* 14(4), 1–115. URL: http://www.digitalhumanities.org/dhq/vol/14/4/000500/000500.html (Accessed: 22 June 2024).

Goldberg, A.E. (1995). *Constructions. A construction grammar approach to argument structure.* Chicago: University of Chicago Press.

Grootendorst, M. (2022). BERTopic. Neural topic modeling with a class-based TF-IDF procedure. *arXiv.* DOI: https://doi.org/10.48550/arXiv.2203.05794 (Accessed: 22 June 2024).

Halter, G., Ballester-Ripoll, R., Flueckiger, B., & Pajarola, R. (2019). VIAN. A Visual Annotation Tool for Film Analysis, *Computer Graphics Forum,* 38(3), 119–129. DOI: https://doi.org/https://doi.org/10.1111/cgf.13676 (Accessed: 22 June 2024).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016* (pp. 770–778). Las Vegas: Institute of Electrical and Electronics Engineers. DOI: https://doi.org/10.1109/CVPR.2016.90 (Accessed: 22 June 2024).

Heftberger, A. (2018). *Digital Humanities and Film Studies. Visualising Dziga Vertov's Work.* Basel: Springer International Publishing.

Hempel, T., Abdelrahman, A.A., & Al-Hamadi, A. (2022). 6d Rotation Representation For Unconstrained Head Pose Estimation. In *Proceedings of the IEEE International Conference on Image Processing 2022* (pp. 2496–2500). Bordeaux: Institute of Electrical and Electronics Engineers. DOI: https://doi.org/10.48550/arXiv.2202.12555 (Accessed: 22 June 2024).

Howanitz, G. (2015). Distant Waching. Ein quantitativer Zugang zu YouTube-Videos. In *DHd 2015. Von Daten zu Erkenntnissen. Book of Abstracts* (pp. 33–38). Graz: Verband Digital Humanities im deutschsprachigen Raum. URL: https://gams.uni-graz.at/o:dhd2015.abstracts-gesamt (Accessed: 22 June 2024).

Hoyt, E., Ponto, K., & Roy, C. (2014). Visualizing and Analyzing the Hollywood Screenplay with ScripThreads, *Digital Humanities Quarterly,* 8(4), 1–57. URL: http://www.digitalhumanities.org/dhqdev/vol/8/4/000190/000190.html (Accessed: 22 June 2024).

Huang, Q., Xiong, Y., Rao, A., Wang, J., & Lin, D. (2020). MovieNet. A Holistic Dataset for Movie Understanding. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Proceedings of the European Conference on Computer Vision 2020* (pp. 709–727). arXiv: Institute of Electrical and Electronics Engineers. DOI: https://doi.org/10.48550/arXiv.2007.10937 (Accessed: 22 June 2024).

Kanzog, K. (1991). *Einführung in die Filmphilologie.* München: Diskurs Film.

Kipp, M. (2014). ANVIL: The Video Annotation Research Tool. In J. Durand, U. Gut & G. Kristoffersen (Eds.), *The Oxford Handbook of Corpus Phonology* (pp. 420–436). Oxford: Oxford University Press.

Korte, H. (2004). *Einführung in die systematische Filmanalyse. Ein Arbeitsbuch.* 3. ed. Berlin: Erich Schmidt Verlag.

Kuang, Z., Sun, H., Li, Z., Yue, X., Lin, T.H., Chen, J., Wei, H., Zhu, Y., Gao, T., Zhang, W., Chen, K., Zhang, W., & Lin, D. (2021). MMOCR. A Comprehensive Toolbox for Text Detection, Recognition and Understanding. In *Proceedings of the 29th ACM International Conference on Multimedia 2021* (pp. 3791–3794). arXiv: Association for Computing Machinery. DOI: https://doi.org/10.48550/arXiv.2108.06543 (Accessed: 22 June 2024).

Kurzhals, K., John, M., Heimerl, F., Kuznecov, P., & Weiskopf, D. (2016). Visual Movie Analytics, *IEEE Transactions on Multimedia,* 18(11), 2149–2160. DOI: https://doi.org/10.1109/TMM.2016.2614184 (Accessed: 22 June 2024).

Liebl, Ch., & Burghardt, M. (2023). Zoetrope. Interactive Feature Exploration in News Videos. In W. Scholger, G. Vogeler, T. Tasovac, A. Baillot, & P. Helling (Eds.), *Digital Humanities 2023. Collaboration as Opportunity* (pp. 432–434). Graz: Alliance of Digital Humanities Organisations. DOI: https://doi.org/10.5281/zenodo.7961822 (Accessed: 22 June 2024).

Liu, C., Feng, L., Liu, G., Wang, H., & Liu, S. (2021). Bottom-up broadcast neural network for music genre classification, *Multimedia Tools and Applications,* 80(5), 7313–7331. DOI: https://doi.org/10.48550/arXiv.1901.08928 (Accessed: 22 June 2024).

Liu, S., Nie, X., & Hamid, R. (2022). Depth-Guided Sparse Structure-from-Motion for Movies and TV Shows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022* (pp. 15980–15989). New Orleans, LA: Institute of Electrical and Electronics Engineers. DOI: https://doi.org/10.48550/arXiv.2204.02509 (Accessed: 22 June 2024).

Loschky, L.C., Larson, A.M., Magliano, J.P., & Smith, T.J. (2015). What would Jaws do? The tyranny of film and the relationship between gaze and higher-level narrative film comprehension, *PloS one,* 10(11), 1–23. DOI: https://doi.org/10.1371/journal.pone.0142474 (Accessed: 22 June 2024).

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

Id. (2020). *Cultural Analytics.* Cambridge, Mass.: MIT Press.

Monaco, J. (2009). *How to Read a Film. Movies, Media and Beyond.* Oxford: Oxford University Press.

Müller-Budack, E., Pustu-Iren, K., & Ewerth, R. (2018). Geolocation Estimation of Photos Using a Hierarchical Model and Scene Classification. In Ferrari, V., Hebert, M., Sminchisescu, C., & Weiss, Y. (Eds.), *Computer Vision. ECCV 2018* (pp. 575–592). Springer, Cham [= *Lecture Notes in Computer Science,* 11216]. DOI: https://doi.org/10.1007/978-3-030-01258-8_35 (Accessed: 22 June 2024).

Müller-Budack, E., Springstein, M., Hakimov, S., Mrutzek, K., & Ewerth, R. (2021). Ontology-driven event type classification in images. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision 2021* (pp. 2928–2938), Waikoloa: Institute of Electrical and Electronics Engineers. DOI: https://doi.org/10.48550/arXiv.2011.04714 (Accessed: 22 June 2024).

Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S., & Ling, H. (2022). Expanding Language-Image Pretrained Models for General Video Recognition. In Avidan, Sh., Brostow, G., Moustapha, C., Farinella, G.M., & Hassner, T. (Eds.), *Proceedings of the European Conference on Computer Vision 2022* (pp. 1–18). Cham: Springer [= *Lecture Notes in Computer Science,* 13664].

Ohm, T., Solà, M.C., Karjus, A. & Schich, M. (2023). Collection Space Navigator. An Interactive Visualization Interface for Multidimensional Datasets, *arXiv.* DOI: https://doi.org/10.48550/arXiv.2305.06809 (Accessed: 22 June 2024).

Prinz, S. (2007). *Movies and meaning. An introduction to film.* 4. ed. Boston: Allyn & Bacon.

Pustu-Iren, K., Sittel, J., Mauer, R., Bulgakowa, O., & Ewerth, R. (2020). Automated Visual Content Analysis for Film Studies. Current Status and Challenges, *Digital Humanities Quarterly,* 14(4), 1–102. URL: http://www.digitalhumanities.org/dhq/vol/14/4/000518/000518.html (Accessed: 22 June 2024).

R-Kernteam. (2016). R. Eine Sprache und Umgebung für statistische Berechnungen [Computersoftware]. *R Foundation for Statistical Computing.* URL: https://www.R-project.org (Accessed: 22 June 2024).

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. In M. Meila & T. Zhang (Eds.), *Proceedings of the International Conference on Machine Learning* (pp. 8748–8763). arXiv. [= *Proceedings of Machine Learning Research,* 139]. DOI: https://doi.org/10.48550/arXiv.2103.00020 (Accessed: 22 June 2024).

Radford, A., Kim, J.W., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the International Conference on Machine Learning 2023* (pp. 28492–28518). Honolulu: International Machine Learning Society [= *Proceedings of Machine Learning Research,* 202]. DOI: https://doi.org/10.48550/arXiv.2212.04356 (Accessed: 22 June 2024).

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., De Mori, R., Bengio, Y. (2021). SpeechBrain. A General-Purpose Speech Toolkit [Preprint]. *arXiv.* DOI: https://doi.org/10.48550/arXiv.2106.04624 (Accessed: 22 June 2024).

Redfern, N. (2022a). Analysing Motion Picture Cutting Rates, *Wide Screen,* 9(1). 1–29. URL: https://widescreenjournal.org/vol-9-no-1-2022-title (Accessed: 22 June 2024).

Id. (2022b). *Computational Film Analysis with R.* Version 0.9.004. Zenodo. DOI: https://doi.org/10.5281/ZENODO.7074521 (Accessed: 22 June 2024).

Ruth, N., Burghardt, M., & Liebl, B. (2023). From Clusters to Graphs. Toward a Scalable Viewing of News Videos. In A. Šeļa, F. Jannidis, & I. Romanowska (Eds.), *Proceedings of the Computational Humanities Research Conference 2023* (pp. 167–177). Paris: Computational Humanities Research. [= *CEUR Workshop Proceedings,* 3558] URL: https://ceur-ws.org/Vol-3558 (Accessed: 22 June 2024).

Ryan, M., & Lenos, M. (2020). *An Introduction to Film Analysis. Technique and Meaning in Narrative Film.* London: Bloomsbury Academic.

Salt, B. (1974). Statistical Style Analysis of Motion Pictures, *Film Quarterly,* 28(1), 13–22. DOI: https://doi.org/10.2307/1211438 (Accessed: 22 June 2024).

Id. (2006). *Moving Into Pictures. More on Film History, Style, and Analysis.* London: Starword Publishing. URL: http://www.starword.com/MovPicFin.pdf (Accessed: 22 June 2024).

Serengil, S.I., & Ozpinar, A. (2021). HyperExtended LightFace. A Facial Attribute Analysis Framework. In *Proceedings of the International Conference on Engineering and Emerging Technologies 2021* (pp. 1–4). Istanbul: IEEE Xplore. DOI: https://doi.org/10.1109/ICEET53442.2021.9659697 (Accessed: 22 June 2024).

Sikov, E. (2010). *Film Studies. An Introduction.* New York City: Columbia University Press.

Sittel, J. (2017). Digital Humanities in der Filmwissenschaft, *MEDIENwissenschaft. Rezensionen. Reviews,* 34(4), 472–489.

Souček, T., & Lokoč, J. (2020). TransNet V2. An effective deep network architecture for fast shot transition detection, *arXiv.* DOI: https://doi.org/10.48550/arXiv.2008.04838 (Accessed: 22 June 2024).

Springstein, M., Stamatakis, M., Plank, M., Sittel, J., Mauer, R., Bulgakowa, O., Ewerth, R., & Müller-Budack, E. (2023). TIB AV-Analytics. Eine webbasierte Plattform für wissenschaftliche Videoanalyse und Filmstudien. In H.-H. Chen & W.-J. Duh (Eds.), *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval 2023* (pp. 3195–3199). New York: Association for Computing Machinery Special Interest Group on Information Retrieval. DOI: https://doi.org/10.1145/3539618.3591820 (Accessed: 22 June 2024).

Stam, R. (2000). *Film Theory. An Introduction.* Malden, Mass.: Blackwell Publishing Limited.

Theiner, J., Müller-Budack, E., & Ewerth, R. (2022). Interpretable Semantic Photo Geolocation. In *Proceedings of the EEE/CVF Winter Conference on Applications of Computer Vision 2022* (pp. 750–760). Waikoloa: Institute of Electrical and Electronics Engineers. DOI: https://doi.org/10.48550/arXiv.2104.14995 (Accessed: 22 June 2024).

Tseng, Ch., Liebl, B., Burghardt, M., & Bateman, J. (2023). FakeNarratives. First Forays in Understanding Narratives of Disinformation in Public and Alternative News Videos. In P. Trilcke, A. Busch, & P. Helling (Eds.), *DHd 2023. Open Humanities Open Culture.* Trier/Luxemburg: Verband Digital Humanities im deutschsprachigen Raum. DOI: https://doi.org/10.5281/zenodo.7715277 (Accessed: 22 June 2024).

Tsivian, Y. (2009). Cinemetrics. Part of the Humanities' Cyberinfrastructure. In M. Ross, M. Grauer & B. Freisleben (Eds.), *Digital Tools in Media Studies. Analysis and Research. An Overview* (pp. 93–100). Bielefeld: Transcript.

Vonderau, P. (2017). Quantitative Werkzeuge. In Hagener, M., & Pantenburg, V. (Eds.), *Handbuch Filmanalyse.* Wiesbaden: Springer VS [= *Springer Reference Geisteswissenschaften*]. DOI: https://doi.org/10.1007/978-3-658-13352-8_28-1 (Accessed: 22 June 2024).

Walkowski, N.-O., & Pause, J. (2018). Everything is Illuminated. Zur numerischen Analyse von Farbigkeit in Filmen, *Zeitschrift für digitale Geisteswissenschaften,* no pag. Wolffenbüttel: Herzog August Bibliothek. DOI: https://doi.org/10.17175/2018_003 (Accessed: 22 June 2024).

Weitin, T. (2017). Skalierbares Lesen, *Zeitschrift für Literaturwissenschaft und Linguistik,* 47, 1–6.

Wevers, M., & Smits, T. (2020). The visual digital turn. Using neural networks to study historical images, *Digital Scholarship in the Humanities,* 35(1), 194–207. DOI: https://doi.org/10.1093/llc/fqy085 (Accessed: 22 June 2024).

Wickham, H. (2016). *ggplot2. Elegant Graphics for Data Analysis.* Berlin/Heidelberg: Springer. DOI: https://doi.org/10.1007/978-3-319-24277-4 (Accessed: 22 June 2024).

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN. A professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation 2006* (pp. 1556–1559). Genoa: ELRA Language Resources Association. URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/153_pdf.pdf (Accessed: 22 June 2024).

Wu, H., Chen, K., Liu, H., Zhuge, M., Li, B., Qiao, R., Shu, X., Gan, B., Xu, L., Ren, B., Xu, M., Zhang, W., Ramachandra, R., Lin, Ch.-W., & Ghanem, B. (2023). NewsNet. A Novel Dataset for Hierarchical Temporal Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023* (pp. 10669–10680). Vancouver: Institute of Electrical and Electronics Engineers. DOI: https://doi.org/10.1109/CVPR52729.2023.01028 (Accessed: 22 June 2024).

Wu, H.-Y., Palù, F., Ranon, R., & Christie, M. (2018). Thinking Like a Director. Film Editing Patterns for Virtual Cinematographic Storytelling, *ACM Transactions* *on Multimedia Computing, Communications, and Applications,* 14(4), 1–22. DOI: https://doi.org/10.1145/3241057 (Accessed: 22 June 2024).

Wu, L., Petroni, F., Josifoski, M., Riedel, S., & Zettlemoyer, L. (2020). Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing 2020* (pp. 6397–6407). arXiv: Association for Computational Lingutistics. DOI: https://doi.org/10.48550/arXiv.1911.03814 (Accessed: 22 June 2024).

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., & Dubnov, S. (2022). Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In *ICASSP 2023. IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 1–5). Rhodes Island: IEEE. DOI: https://doi.org/10.1109/ICASSP49357.2023.10095969 (Accessed: 22 June 2024).

Zacks, J.M. (2010). Wie wir unsere Erfahrungen zu Ereignissen organisieren, *Psychological Science Agenda,* 24(4).

Zhang, H., Yuan, T., Chen, J., Li, X., Zheng, R., Huang, Y., Chen, X., Gong, E., Chen, Z., Hu, X., Yu, D., Ma, Y., & Huang, L. (2022). PaddleSpeech. An Easy-to-Use All-in-One Speech Toolkit. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics. Human Language Technologies. System Demonstrations* (pp. 114–123). Seattle: Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2022.naacl-demo.12 (Accessed: 22 June 2024).

Zhang, H., Li, X., & Bing, L. (2023). Video-LLaMA. An Instruction-tuned Audio-Visual Language Model for Video Understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. System Demonstrations* (pp. 543–553). Singapur: Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2023.emnlp-demo.49 (Accessed: 22 June 2024).

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places. A 10 Million Image Database for Scene Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 40(6), 1452–1464. DOI: https://doi.org/10.1109/TPAMI.2017.2723009 (Accessed: 22 June 2024).

Figure Credits

Fig. 1–4 are self-created screenshots from the authors' work with TIB AV-A (Fig. 1) and the R package *ggplot* (Fig. 2–4). They are all published here for the first time.