Computergestützte Werkzeuge und Methoden für die Film- und Videoanalyse

Manuel Burghardt[a], John Bateman[b], Eric Müller-Budack[c] und 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, [d] https://orcid.org/0000-0003-0918-6297

Abstract Dieses Kapitel gibt einen Überblick über computergestützte Werkzeuge und Methoden für die Film- und Videoanalyse. Nach einem kurzen historischen Überblick zur Entwicklung bestehender Methoden in diesem Bereich werden sodann die theoretischen Grundlagen für empirische Videoanalysen gelegt. Da sich das Kapitel auf den Einsatz aktueller Deep-Learning-Methoden konzentriert, gibt es auch einen Überblick über die Arten von Informationen, die mit diesen Methoden extrahiert werden können. Darüber hinaus wird ein einfach zu bedienendes Tool für die Videoanalyse vorgestellt, genannt TIB AV-Analytics (TIB AV-A). Es wird demonstriert, wie TIB AV-A eingesetzt werden kann, um die Erforschung narrativer Muster in der beliebten Fernsehserie *Game of Thrones* zu unterstützen. Abschließend wird der aktuelle Stand der verfügbaren Werkzeuge und Methoden für die computergestützte Videoanalyse zusammengefasst und künftige Herausforderungen skizziert.*

Keywords Computergestützte Filmanalyse, computergestützte Videoanalyse, distant viewing

1. Eine kurze Geschichte der computergestützten Methoden in der Film- und Videowissenschaft

Das Feld der Digital Humanities (DH) hat traditionell einen starken Fokus auf textuelles Material, dessen Ursprünge auf Roberto Busas *Index Thomisticus* zurückgeführt werden. In den letzten Jahren hat jedoch das Interesse an Film und Video innerhalb der DH stark zugenommen (Burghardt & Wolff 2016; Sittel 2017; Heftberger 2018; Burghardt et al. 2020; Arnold & Tilton 2022), was zur Gründung spezieller Interessengruppen sowohl auf nationaler[1] als auch auf internationaler[2] Ebene führte. Ange-

* Dieses Kapitel wurde inkl. fremdsprachiger Zitate von der Redaktion aus dem Englischen übersetzt.

1 S. DHd AG Film & Video: https://dig-hum.de/ag-film-und-video, zuletzt aufgerufen am 22.06.2024.

2 S. ADHO Special Interest Group AudioVisuelles Material in den Digital Humanities: https://avindhsig.wordpress.com, zuletzt aufgerufen am 22.06.2024.

Manuel Burghardt, John Bateman, Eric Müller-Budack und Ralph Ewerth: Computergestützte Werkzeuge und Methoden für die Film- und Videoanalyse, in: Christopher A. Nunn und Frederike van Oorschot (Hrsg.): Kompendium Computational Theology 1. Forschungspraktiken in den Digital Humanities, Heidelberg: heiBOOKS, 2024, S. 161–189. DOI: https://doi.org/10.11588/heibooks.1459.c21909

sichts des hochgradig interdisziplinären Charakters der DH haben sich verschiedene Perspektiven auf die Analyse von Film und Video herausgebildet. Burghardt et al. (2020) identifizieren drei Hauptperspektiven: (1) Eine *infrastrukturelle Perspektive*, die GLAM-Institutionen (Galleries, Libraries, Archives, Museums) und Filmarchive umfasst, (2) eine *Medienperspektive,* die sich mit digitalen Begegnungen in der Film- und Medienwissenschaft befasst, und (3) eine *computergestützte Perspektive,* die sich auf *multimediales Information Retrieval* und *multimodale Informationsextraktion* konzentriert. Da immer mehr Videomaterial in digitaler Form zur Verfügung steht, hat die computergestützte Perspektive erheblich an Bedeutung gewonnen, was zur Entstehung von Begriffen wie „distant viewing" (Arnold & Tilton 2019), „distant watching" (Howanitz 2015) und „deep watching" (Bermeitinger et al. 2019) geführt hat. Wevers & Smits (2020) schlagen sogar einen *visual digital turn* vor, der durch die Möglichkeiten von Deep-Learning-Techniken vorangetrieben wird.

Während Ansätze wie distant viewing u. ä. in jüngster Zeit an Aufmerksamkeit gewonnen haben, finden sich frühe Beispiele quantitativer Filmstudien mit einem Schwerpunkt auf Einstellungsanalysen bereits bei Barry Salt (1974; 2006) und Yuri Tsivians Datenbank *Cinemetrics* (2009). Darüber hinaus wurden weitere quantitative Ansätze zur stilistischen und formalen Analyse von Filmmerkmalen erforscht. Dazu gehören zum Beispiel die Sprachanalyse (Hoyt et al. 2014; Byszuk 2020; Bednarek 2023) und die Farbanalyse (Burghardt et al. 2018; Pause & Walkowski 2018; Flueckiger & Halter 2020). Ein anderer Zweig der quantitativen Analyse großer Videokorpora hat seinen Schwerpunkt in der Informationsvisualisierung. Manovichs (2013) Projekt *Visualizing Vertov* stellt ein frühes Beispiel dar, das sich stark auf Visualisierungen stützt, um Muster in umfangreichen Bild- und Videosammlungen sichtbar zu machen. Zusätzlich zu solchen visuellen Analyseansätzen gibt es eine breite Palette von Tools für die Annotation und Analyse von Videos und Filmen. Beispiele für diese Tools sind ELAN (Wittenburg et al. 2006), Videana (Ewerth et al. 2009), ANVIL (Kipp 2014) und VIAN (Halter et al. 2019). Für einen umfassenden Überblick über diese Tools sowie ihre spezifischen Merkmale und Funktionen empfehlen wir das Übersichtspapier von Pustu-Iren et al. (2020). Darüber hinaus gibt es neuere Tools, die über die reine Analyse visueller Aspekte hinausgehen und sowohl Sprache (gesprochen und geschrieben) als auch Audio (Musik und Geräusche) einbeziehen. Beispiele für solche Tools sind Zoetrope (Tseng et al. 2023; Liebl & Burghardt 2023) und TIB AV-A (Springstein et al. 2023).

Mit diesem Kapitel wollen wir eine Einführung in die computergestützte, empirische Analyse von Filmen und Videos geben. In Abschnitt 2 legen wir einige theoretische Grundlagen für solche empirischen Analysen. Abschnitt 3 bietet einen Überblick über die Arten von Informationen, die mit Methoden des Deep Learning extrahiert werden können. Da die Einrichtung solcher Verfahren ohne fortgeschrittene technische Kenntnisse eine Herausforderung sein kann, wird auch ein einfach zu bedienendes Tool namens TIB AV-Analytics (TIB AV-A, Abschnitt 4) vorgestellt. Darüber hinaus wird in Abschnitt 5 eine Fallstudie präsentiert, die zeigt, wie TIB AV-A für die

Erforschung von Erzählmustern in der beliebten Fernsehserie *Game of Thrones* eingesetzt werden kann. Abschnitt 6 fasst abschließend den aktuellen Stand der verfügbaren Werkzeuge und Methoden für die computergestützte Videoanalyse zusammen und skizziert einige Herausforderungen, die vor uns liegen.

2. Theoretische Grundlagen der empirischen Filmanalyse

Bei der Untersuchung von Filmen wird seit langem eine Kombination aus qualitativen und quantitativen Methoden angewandt (Korte 2004). Während qualitative Ansätze bei der Analyse von Filmen weit verbreitet sind, wie Studien von Stam (2000), Prince (2007), Sikov (2010), Ryan & Lenos (2020) und viele andere zeigen, betonen DH-Ansätze in der Filmwissenschaft eher quantitative und empirische Aspekte. In diesem Abschnitt sollen einige theoretische Grundlagen geschaffen werden, die empirische Filmstudien weiter kontextualisieren und solche Bemühungen mit breiteren analytischen Anliegen in Verbindung bringen.

Eine Vielzahl *externer* Perspektiven auf Film ist auch von großer Relevanz für die Beschäftigung mit Film innerhalb der DH. Diese reichen von Archivierung, historischen Studien, Produktionsstudien, Untersuchungen der Auswirkungen von sich entwickelnden Anzeigetechnologien, der Bereitstellung von Informationen aus audiovisuellen Daten, Rezeptionsstudien (von psychologischen Studien bis hin zu Rezensionen und Kritiken) bis hin zu Kulturstudien über das Kino als Institution. In allen Fällen hilft es zum Verständnis des Mediums, wenn man einen engeren analytischen Zugriff auf das *Innere* des Films erhält, d. h. Filme als Artefakte mit spezifischen Designs für ästhetische oder andere Zwecke versteht. Eine solche analytische Haltung lässt sich am besten mit Hilfe empirischer Studien erreichen, bei denen die Eigenschaften von Filmen nach und nach aufgedeckt werden, um immer stärkere Verallgemeinerungen über ihre Funktionsweise abzuleiten. DH-Ansätze zum Film suchen daher typischerweise nach Forschungsstrategien, die sowohl quantitative Studien der Verteilungen und Muster von messbaren Filmeigenschaften als auch qualitative, eher hermeneutische Interpretationen dieser Verteilungen und Muster umfassen (Flückiger 2011; Heftberger 2018).

Die zentrale Herausforderung in diesem Zusammenhang ist die Frage, wie man Zugang zu den Details der hochkomplexen audiovisuellen Filmartefakte erhält, die für die Interpretation und Analyse relevant sind. Da viele der im Zusammenhang mit Film aufgeworfenen Fragen interpretativer und hermeneutischer Natur sind, ist es keineswegs selbstverständlich, dass und wie quantitative Ansätze solche Anliegen unterstützen können. Dies ist freilich eine sehr allgemeine philosophische Frage, die für viele Zweige der DH gilt. Im Falle des Films wurden die frühen quantitativen Ansätze aus der Stilometrie in den Literaturwissenschaften angeregt (vgl. das Kapitel von F. Jannidis in diesem Band). Diese waren jedoch auf manuelle Methoden be-

schränkt, die sowohl den Umfang der Studien als auch die Art der zu berücksichtigenden Merkmale einschränkten. Das umfangreichste Programm dieser Art, das viele Jahre lang von Barry Salt (1974; 2007) durchgeführt wurde, zählte die Einstellungslängen und -größen für ausgewählte Teile von Filmsammlungen aus verschiedenen Epochen und von unterschiedlichen Regisseuren. Die Arbeiten in dieser Tradition werden fortgesetzt und haben eine Vielzahl historischer und regionaler Entwicklungen aufgezeigt; Cutting & Candan (2015) berichten z. B. über eine wiederum überwiegend manuelle Analyse der Einstellungslängen von 9 400 englischsprachigen und 1 550 nicht-englischsprachigen Filmen, die zwischen 1912 und 2013 veröffentlicht wurden. Ähnlich weitreichende historische Veränderungen wurden für Helligkeit, Farbe und Szenenübergänge berichtet (z. B. Cutting et al. 2011a; b). Redfern (2022b) stellt eine Vielzahl von praktisch anwendbaren R-Skripten (R Core Team 2016) vor, die diese Art von Studien unterstützen.

Heftberger (2018) verfolgt einen anderen Ansatz, indem sie sich auf direkte Visualisierungen stützt, die aus Standbildern von Filmen bestehen und sich insbesondere auf die Werke von Dziga Vertov konzentrieren. Direkte Visualisierungen von wahrnehmbaren Merkmalen wie Helligkeit, Farbe usw. wurden von mehreren DH-Forscher*innen untersucht, da die Erzeugung solcher visuellen Darstellungen relativ einfach ist. Es ist jedoch fraglich, inwieweit das bloße Verlassen auf die (typischerweise) visuelle Wahrnehmung eine wirksame Methode ist, um bedeutsame Muster aufzudecken. In vielerlei Hinsicht ist dies symptomatisch für die derzeitige sehr explorative Phase der DH-Filmstudien, in der nur das für die Analyse verwendet wird, was technologisch machbar ist. Bislang sind umfassendere Ansätze, die sich mit einem breiteren Spektrum von Filmphänomenen befassen, noch begrenzt. Bakels et al. (2020) berichten zum Beispiel von sehr detaillierten Analysen von Filmen auf mehreren Ebenen, die speziell auf Fragen der audiovisuellen Konstruktion von Affekten in Filmen abzielen und sich dabei auf die AdA-Ontologie filmanalytischer Begriffe stützen.[3] Allerdings handelt es sich auch hier noch weitgehend um manuelle Verfahren, wenngleich nun sukzessive auch automatisierte digitale Techniken hinzukommen.

Es liegt auf der Hand, dass empirische Ansätze erheblich von technologischer Unterstützung profitieren können, sodass die zeitaufwendige und fehleranfällige manuelle Analyse auf ein Minimum reduziert werden kann, aber es bleiben grundsätzliche Fragen hinsichtlich des allgemeinen Nutzens solcher Ansätze für den Film. Für die weitere Entwicklung wird es wichtig sein, die Möglichkeiten rechnergestützter Analysewerkzeuge stärker auf die Forschungsfragen der Filmwissenschaft zu beziehen. Hier gibt es bislang noch eine Lücke, die es zu schließen gilt. Heftberger begründet dies damit, dass Vertov selbst, wie sie anmerkt, bei der Konstruktion seiner Filme auf formale Gestaltungsmerkmale geachtet hat und somit die formale Analyse

3 S. https://projectada.github.io/ontology, zuletzt aufgerufen am 22.06.2024.

durchaus gerechtfertigt ist. Es ist jedoch nicht klar, ob dies als allgemeine Leitlinie für die empirische Analyse von Filmen mit Hilfe von computergestützten und anderen quantitativen Methoden gelten kann.

Eine der Methoden, mit denen derzeit allgemeinere Orientierungen entwickelt werden, ist stark geprägt von einer multimedialen *Information-Retrieval*-Perspektive auf die computergestützte Filmanalyse. Kurzhals et al. (2016) versuchen beispielsweise, die *visuelle Filmanalyse* auf eine semantischere Ebene zu heben, indem sie eine Vielzahl von Informationsquellen (einschließlich Drehbüchern und Untertiteln) kombinieren, um die vier grundlegenden Fragen der Zusammenfassung zu beantworten: *wer, was, wo* und *wann.* In dieser Architektur wird eine breite Palette automatischer Verarbeitungstechniken kombiniert, um Beschreibungen von Szenen und ihren Ereignissen zu liefern, wodurch einige der traditionellen Ansätze zur Filmanalyse in Form von detaillierten *shot-by-shot*-Beschreibungen, die als Film- oder Sequenz-Protokolle bekannt sind, aus einer ganz anderen Perspektive aufgegriffen werden (vgl. Kanzog 1991, 136–151; 163–183). Die Erstellung solcher Protokolle ist jedoch äußerst zeitaufwendig, sodass jeder Beitrag zu deren automatischer Erstellung einen nützlichen Fortschritt darstellt. In den kommenden Jahren werden noch viele solcher Möglichkeiten entstehen und einige wichtige Entwicklungen in diese Richtung werden weiter unten besprochen. Nichtsdestotrotz und speziell im Hinblick auf den Film als ästhetisches Artefakt fällt auf, dass die vier oben genannten Fragen Folgendes auslassen: eine Frage, die für die Filmanalyse entscheidend ist, nämlich die Frage nach dem *Wie.* Es reicht oft nicht aus, nur die bloße Handlungsstruktur einer Erzählung zu beschreiben. Bei Filmen sind wir oft ebenso an der Art und Weise des Erzählens interessiert, denn das ist es, was den Film als kulturell wirksame Kommunikationsform ausmacht. Die Einbeziehung solcher Informationen war einer der Hauptgründe, warum traditionelle Filmprotokolle so arbeitsintensiv zu produzieren waren.

Diese führt uns direkt zu einigen grundlegenden Fragen bezüglich der stärker interpretativen Natur der Filmanalyse, welche oben bereits erwähnt wurden und die die DH noch in den Griff bekommen muss. Da es in der Regel keineswegs einfach ist, formale Merkmale von Filmen mit Interpretationen in Verbindung zu bringen, sind Strategien zur weiteren Abstraktion erforderlich, um die Kluft zu überbrücken. Vonderau (2017) gibt z. B. eine nützliche Zusammenfassung einiger der Auseinandersetzungen zwischen eher digital orientierten Ansätzen zum Film als Daten einerseits und den traditionellen Anliegen der Filmwissenschaft andererseits. Diese in den DH neu aufgeworfenen Fragen haben aber auch eine längere Geschichte als wiederkehrende Kritik von Filmwissenschaftlern an rein quantitativen Ansätzen im Allgemeinen. Schon früh stellten David Bordwell und Kolleg*innen fest, dass es zwar möglich ist, allgemeine statistische Normen im Stile von Barry Salt zu berechnen, dass aber solche Abstraktionen wenig bedeuten, wenn man kein Konzept für die Bandbreite der zu einem bestimmten Zeitpunkt vorherrschenden paradigmatischen Wahl hat (Bordwell et al. 1985, 60). Darüber hinaus ist es notwendig, sich mit folgenden

Aspekten einer wesentlichen Indirektheit im Prozess der Bedeutungszuschreibung im Medium Film zu befassen, die bei quantitativen Ansätzen leicht übersehen werden:

> Manchmal sind wir versucht, Winkeln, Abständen und anderen Qualitäten der Bildgestaltung absolute Bedeutungen zuzuweisen. […] Die Analyse des Films als Kunst wäre viel einfacher, wenn technische Qualitäten automatisch solche festen Bedeutungen hätten, aber die einzelnen Filme würden dadurch viel von ihrer Einzigartigkeit und ihrem Reichtum verlieren. Tatsache ist, dass Framings keine absoluten oder allgemeinen Bedeutungen haben. (Bordwell & Thompson 2008, 192)

Nur innerhalb bestimmter Kontrastsysteme können überhaupt spezifische Bedeutungen zugewiesen werden (Branigan 1984, 29). Die Untersuchung dieses Aspekts der filmischen Bedeutungsgebung wird in den aktuellen computergestützten Werkzeugen nur über interaktive Schnittstellen unterstützt, da die computergestützten Verfahren selbst noch nicht in der Lage sind, verlässliche Hypothesen zur Interpretation zu liefern. Es ist daher, wie in den folgenden Abschnitten sichtbar wird, sinnvoll, einen schnellen Zugriff auf automatisch erfasste formale Merkmale von Filmen zu ermöglichen, die Interpretationen solch komplexer Merkmalsbündel bleibt jedoch menschlichen Analytiker*innen überlassen, indem sie entsprechende Abfragen über Kombinationen von automatisch markierten Kategorien manuell formulieren (vgl. z. B. Kurzhals et al. 2016).

Die direktere Unterstützung von Interpretationsaufgaben wird zweifellos ein wichtiger Entwicklungsbereich für die Zukunft sein. Hierfür wird es nützlich sein, robustere theoretische Rahmenwerke zu konstruieren, die sich mehr als bisher in den DH auf explizite semiotische Grundlagen stützen. Solche Grundlagen müssen sich an zeitgenössischen Auffassungen von Semiotik orientieren, die einerseits quantitative und qualitative Beschreibungsformen gleichermaßen unterstützen und andererseits in der Lage sind, sowohl die Anerkennung formaler technischer Merkmale als auch den gesamten Prozess der hermeneutischen Interpretation im Kontext zu umfassen. Nur dann können Studien ein Gleichgewicht zu den derzeit vorherrschenden, rein datengesteuerten Bottom-up-Ansätzen herstellen (Redfern 2022b).

Eine solche Darstellung der Semiotik, die sowohl mit den DH als auch mit der Analyse von Filmen in Verbindung gebracht wurde, wird in Bateman et al. (2017) dargelegt. In diesem Ansatz werden die Ausdrucksmittel eines Mediums, zu denen beim Film Montage, Beleuchtung, Farbe, Musik und vieles mehr gehören, jeweils auf drei verschiedenen Abstraktionsebenen charakterisiert: *Material* (zur Unterstützung von Messungen), *Form* (zur Unterstützung expliziter Darstellungen paradigmatischer Systeme von Kontrasten) und *Diskurs* (zur Unterstützung der Interpretation im Kontext). Die letztgenannte Beschreibungsebene geht über die Möglichkeiten der eher strukturellen Semiotik hinaus, die sich in den 1960er Jahren durchsetzte. Die verschiedenen Abstraktionsebenen sind für die DH von unmittelbarer Bedeutung, da sie

verschiedene Klassen von Annotationen motivieren, die zur Beschreibung beliebiger analysierter Artefakte eingesetzt werden können und somit zur Organisation größerer Sammlungen dienen (Bateman 2022).

Die multiperspektivische Sichtweise ist nun auch für die Einbindung der neuen Generation von Deep-Learning-basierten Computertechniken in einem kohärenten Gesamtrahmen unerlässlich. Solche Techniken sind nicht mehr auf einzelne Abstraktionsebenen beschränkt und liefern nützliche Ergebnisse, die von formalen Merkmalen eines Films auf niedriger Ebene bis hin zu direkten Beschreibungen semantischer Inhalte reichen. Semiotisch gesehen funktionieren diese Komponenten daher ähnlich wie der linguistische Begriff der Konstruktionen, die typischerweise Informationen aus verschiedenen Abstraktionsebenen kombinieren, um wiederverwendbare Bausteine für die Kommunikation anzubieten (Goldberg 1995). Speziell filmische Konstruktionen, die oft als filmische Idiome bezeichnet werden, werden nun auch formal behandelt (Wu et al. 2018). Der logische nächste Schritt ist daher, diese Beschreibungen mit automatischen Analysekomponenten zur Vorhersage und Erkennung zu kombinieren und sowohl Visualisierungen als auch statistische Auswertungen zu unterstützen, die auf den erreichten höheren Abstraktionsebenen aufbauen. Erste Schritte in diese Richtung werden im Folgenden vorgeschlagen.

3. Eine kurze Einführung in multimodale Informationsextraktionssysteme

Videos umfassen mehrere ausdrucksstarke Modalitäten wie Bild, Audio (einschließlich Sprache) und Text (in Videobildern eingeblendeter Text). Zur Analyse einzelner Videos sowie auch größerer Korpora werden Informationen aus allen Modalitäten benötigt. Da die manuelle Analyse von Filmen eine sehr zeitaufwendige Aufgabe ist, besteht ein großer Bedarf an automatischen Mustererkennungs- und Multimedia-Retrieval-Methoden zur Unterstützung von DH-Forscher*innen. In den letzten Jahren wurden dank Deep-Learning-Modellen und der Verfügbarkeit großer Datensätze für das Training enorme Fortschritte in Bereichen der Informatik wie Computer Vision, Audioanalyse und Verarbeitung natürlicher Sprache erzielt. Dieser Abschnitt gibt einen Überblick über Ansätze zur Informationsextraktion aus Bildern, Audiodaten und Texten für die Filmanalyse, wobei zu beachten ist, dass es sich nur um eine Auswahl von denjenigen Ansätzen handelt, die wir aufgrund unserer Zusammenarbeit mit Forschenden aus den DH (d. h. Filmanalyst*innen, Semiotiker*innen, Medien- und Kommunikationswissenschaftler*innen) für die Videoanalyse als besonders relevant erachten. Darüber hinaus gibt es weitere multimodale Ansätze, die die Potentiale der hier aufgeführten Methoden kombinieren.

Computer Vision: Für die Filmanalyse sind verschiedene Aspekte visueller Informationen wichtig, die von einfachen Merkmalen (z. B. Farbe, Helligkeit) über Kameraeinstellungen (z. B. Einstellungsgröße, Kamerabewegung) bis hin zu komplexeren Informationen (z. B. Handlungen, Orte, Personen) reichen. Es gibt viele Bibliotheken (z. B. *scikit-learn*[4]), um Low-Level-Merkmale wie *Helligkeit, Farbe* und *Kontrast* zu extrahieren. Für die meisten anderen Computer-Vision-Aufgaben werden in der Regel Deep-Learning-Modelle wie *convolutional neural networks* (z. B. He et al. 2016) oder Transformer-Modelle (z. B. Radford et al. 2021) verwendet. Methoden zur *zeitlichen Videosegmentierung* sind ein wesentlicher Schritt zur Strukturierung eines Videos und können in Bezug auf die *Erkennung von Einstellungen* (z. B. Souček & Lokoč 2020), *Szenen* und *Themengrenzen* (z. B. Wu et al. 2023) kategorisiert werden. Auch relevante Informationen hinsichtlich der Kameraeinstellung wie beispielsweise Aufnahmegröße, Kamerabewegung, -winkel und -orientierung können mit Deep-Learning-Verfahren (z. B. Huang et al. 2020; Liu et al. 2022) geschätzt werden. Ansätze zur *optical character recognition* (z. B. Kuang et al. 2021) erkennen automatisch überlagerten Text in Bildern, der mit Ansätzen zur Verarbeitung natürlicher Sprache weiter analysiert werden kann (siehe unten). Für die Analyse von Bildinhalten gibt es verschiedene Deep-Learning-Ansätze. Insbesondere die Identifizierung von *Personen* (z. B. Deng et al. 2020), *Gesichtsmerkmalen* (z. B. Emotionen, Kopfhaltung, Geschlecht; Hempel et al. 2022; Serengil & Ozpinar 2021) und anderen *Konzepten* (z. B. Tiere, Autos, Objekte; Radford et al. 2021) wurde von der Computer-Vision-Community breit erforscht. Darüber hinaus wurden Ansätze zur Identifizierung von *Ortskategorien* (z. B. Kirche, Markt, Restaurant; Zhou et al. 2018), *geografischen Orten* (z. B. Müller-Budack et al. 2018; Theiner et al. 2022) und *Ereignissen* (z. B. Proteste, Wahlen, Naturkatastrophen; Müller-Budack et al. 2021) vorgestellt, die zur Kategorisierung und Charakterisierung von Filmsegmenten verwendet werden können. Während Deep-Learning-Modelle häufig explizit für solche Aufgaben optimiert werden, indem gelabelte Trainingsdaten verwendet werden, wurden neuere Bild-Sprach-Modelle wie *CLIP* (*Contrastive Language-Image Pretraining;* Radford et al. 2021) mit Hunderten von Millionen Bild-Text-Paaren trainiert, um implizit visuelle Konzepte zu lernen. Diese Modelle können für viele Aufgaben eingesetzt werden, da sie die Ähnlichkeit *beliebiger Konzepte* (z. B. Objekte, Wetter, Beruf) mit einem Bild auf der Grundlage einer textuellen Beschreibung messen können. Kürzlich wurden neue multimodale *large language models* (z. B. Alayrac et al. 2022; Dai et al. 2023) entwickelt. Sie kombinieren die Fähigkeiten dieser Ansätze mit großen Sprachmodellen wie dem GPT-4[5] (*Generative Pre-training Transformer* 4) von OpenAI und erzielen beeindruckende Ergebnisse für viele Anwendungen, einschließlich Film- und Videoanalyse (Zhang et al. 2023). Während sich die meisten der genannten Ansätze auf Einzelbilder konzentrieren und auf jedes einzelne Videobild angewendet werden müssen, berücksichtigen

4 S. https://scikit-learn.org, zuletzt aufgerufen am 22.06.2024.

5 S. https://openai.com/gpt-4, zuletzt aufgerufen am 22.06.2024.

Methoden zur Videoklassifikation auch den zeitlichen Kontext von Bildsequenzen (z. B. Ni et al. 2022) für weitere Anwendungen wie die *Erkennung von Handlungen* (z. B. Laufen, Sprechen).

Audio-Analyse: Grundlegende Analyseschritte für Audio betreffen *Low-Level-Merkmale* wie *Amplitude, Lautstärke* und *Spektrogramm* (z. B. unter Verwendung der *librosa*-Bibliothek für *Python*), welche die Lautstärkeänderungen, (rhythmische) Muster, Musik und andere Soundeffekte anzeigen können. Die *Transkription der gesprochenen Sprache* ist eine weitere sehr wichtige Aufgabe für die Filmanalyse. Kürzlich wurden neuronale Transformer-Architekturen für die automatische Spracherkennung eingeführt (z. B. *Whisper;* Radford et al. 2023), die in vielen Sprachen beeindruckende Ergebnisse erzielen.[6] Automatisch extrahierte Transkripte ermöglichen eine eingehende Analyse von Sprache mit Hilfe von Werkzeugen der natürlichen Sprachverarbeitung (siehe unten). Methoden zur *Sprecher*innen-Erkennung* (z. B. Bredin & Laurent 2021) können das Sprachtranskript weiter verfeinern, indem sie die Identität der jeweiligen Sprecher*innen zuordnen, um z. B. die gesprochene Sprache für alle Sprecher*innen einzeln zu analysieren oder um Gesprächsformen (z. B. Monolog, Dialog) in einem Film zu finden. Sie dient auch als Grundlage für die *Identifizierung von Stimmmerkmalen* wie Geschlecht (z. B. Baevski et al. 2020) und Emotionen (z. B. Ravanelli et al. 2021). Neben der Analyse von Sprache haben sich Forscher*innen auch auf die *Erkennung und Klassifizierung von Musik* (z. B. Liu et al. 2021) sowie auf eine allgemeinere *Audioklassifizierung* (z. B. Wu et al. 2022) konzentriert. Motiviert durch CLIP (siehe oben) wurde CLAP (*Contrastive Language-Audio Pretraining;* Wu et al. 2022) mit mehreren hunderttausend Audio- und Textpaaren trainiert, um die Klassifizierung *beliebiger* Audiokonzepte (z. B. Klangereignisse wie *Sirenengeheul* oder *Regen*) auf der Grundlage von textuellen Prompts zu ermöglichen.[7]

Verarbeitung natürlicher Sprache: Wie bereits erwähnt, ermöglichen Methoden zur optischen Zeichenerkennung aus Bildern (Videoframes) und zur automatischen Spracherkennung aus Audiodaten die Extraktion von Textinformationen aus Videos auf der Grundlage von überlagertem Text und Sprache. Methoden aus der Verarbeitung natürlicher Sprache eröffnen viele Perspektiven für die weitere Arbeit mit solchen Sprachdaten. So kann z. B. das *Part-of-Speech-Tagging* (z. B. *spaCy*[8]) zur Syntaxanalyse eingesetzt werden, um die grammatikalische Struktur eines Satzes besser zu verstehen. *Named Entity Recognition* und *Disambiguation* (z. B. *spaCy,* Wu et al. 2020) können automatisch Erwähnungen von Personen, Orten und Ereignissen erkennen,

6 S. https://github.com/openai/whisper#available-models-and-languages, zuletzt aufgerufen am 22.06.2024.

7 Vgl. den Beitrag von Ch. Weiß in diesem Band.

8 S. https://spacy.io, zuletzt aufgerufen am 22.06.2024.

die in Videos und Filmen eine wichtige Rolle spielen.[9] Darüber hinaus gibt es zahlreiche Ansätze für die Klassifizierung von *Themen* (z. B. Grootendorst 2022) und *Stimmungen* (z. B. Devlin et al. 2019), die Aufschluss über die Gesamthandlung sowie den emotionalen Ton und die Dynamik der Figuren im Film geben können.[10] In jüngster Zeit wurden auch große Sprachmodelle wie das GPT-4 von OpenAI in großem Umfang für eine Vielzahl der oben genannten Aufgaben und darüber hinaus eingesetzt.

4. Videoanalyse mit dem TIB AV-Analytics (TIB AV-A) Tool

Die Implementierung der im vorherigen Abschnitt vorgestellten Deep-Learning-Techniken kann erhebliche technische Herausforderungen mit sich bringen. Als Zwischenschritt wurde eine Reihe von Toolkits[11] vorgeschlagen, die eine grundlegende Abstraktionsebene bieten, aber dennoch fortgeschrittenes technisches Wissen und Datenkenntnisse erfordern. Um jedoch die Vorteile groß angelegter Mustererkennungs- und Multimedia-Retrieval-Methoden (siehe Abschnitt 3) einer größeren Forschungscommunity zugänglich zu machen, die mit audiovisuellem Material arbeitet, ist ein einfach zu bedienendes Werkzeug mit einer grafischen Benutzeroberfläche wünschenswert. Dies ist die Hauptmotivation für die TIB AV-Analytics-Plattform (TIB AV-A)[12], die derzeit vom TIB – Leibniz-Informationszentrum für Technik und Naturwissenschaften in Zusammenarbeit mit Filmwissenschaftler*innen der Universität Mainz entwickelt wird.

TIB AV-A ist eine webbasierte Plattform zur systematischen Film- und Videoanalyse (ein Screenshot ist in Abb. 1 zu sehen). Die Plattform nutzt moderne Webtechnologien und eine Plugin-Struktur, um die Integration neuer Plugins für Entwickler*innen und Forschende zu vereinfachen und TIB AV-A auf dem aktuellen Stand der Technik zu halten. Wir verwenden Container (z. B. *Docker*[13]) zur Virtualisierung für eine einfache Einrichtung und zur Verwaltung von Software-Abhängigkeiten sowie einen *Inference Server* (aktuell *Ray*[14]) für eine stabile Bereitstellung. Um die Interoperabilität mit anderen Videoanalysetools zu gewährleisten, bietet TIB AV-A eine Anwendungsprogrammierschnittstelle (API) und den Import und Export von Ergebnissen in gängigen Datenformaten sowie für das weit verbreitete ELAN-Videoanno-

9 Vgl. den Beitrag von E. Gius in diesem Band.
10 Vgl. die Beiträge von M. Althage und R. Sprugnoli in diesem Band.
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. Beide Adressen wurden zuletzt am 22.06.2024 aufgerufen.
12 S. https://service.tib.eu/tibava, zuletzt aufgerufen am 22.06.2024.
13 S. https://www.docker.com, zuletzt aufgerufen am 22.06.2024.
14 S. https://www.ray.io, zuletzt aufgerufen am 22.06.2024.

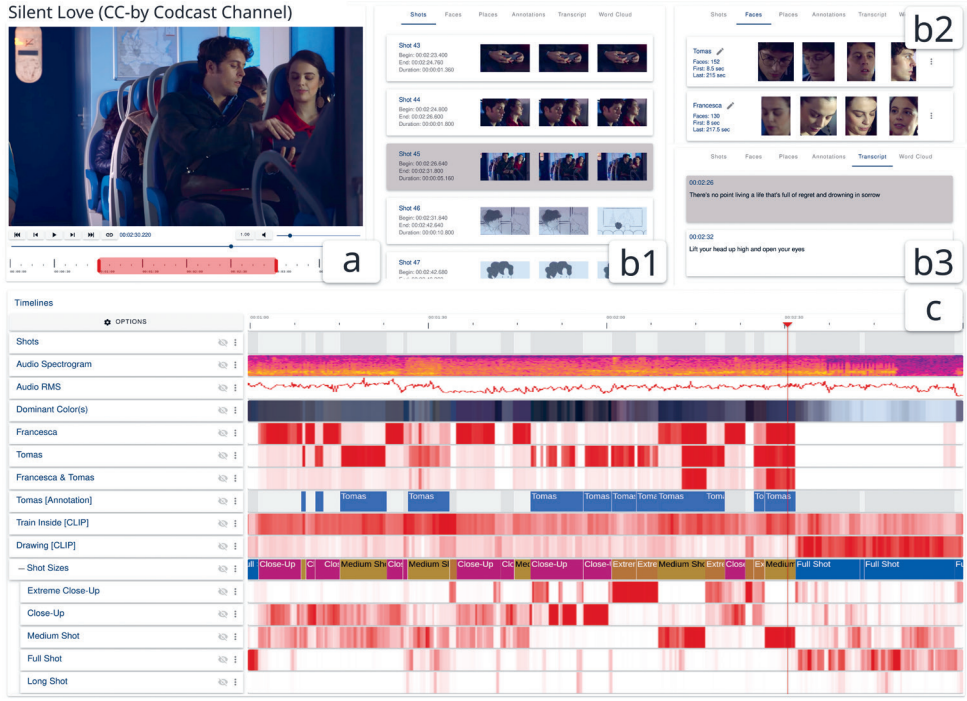

Abb. 1 Benutzungsoberfläche von TIB AV-A für den Kurzfilm *Silent Love* (CC-by Codcast Channel, Originalvideo: https://www.youtube.com/watch?v=KuuEs0oVVS8). Sie enthält einen Videoplayer (a), eine Übersicht der erkannten Einstellungen (b1), Personen (b2) und das Sprachtranskript (b3). Die Zeitleisten (c) können kategorische (z. B. „Francesca") und numerische Werte (z. B. *Zeichnung* [CLIP]) anzeigen. Zeitleisten mit numerischen Werten zeigen z. B. die Wahrscheinlichkeit an, ob ein Konzept in einem Video dargestellt wird. Nutzer*innen können die Art der Visualisierung (Liniendiagramm, Farbdiagramm) und die Farbe (hier: von weiß [unwahrscheinlich] bis rot [wahrscheinlich]) auswählen.

tationstool (Wittenburg et al. 2006). Der Quellcode ist öffentlich zugänglich.[15] Weitere technische Details werden in Springstein et al. (2023) beschrieben.

Im Gegensatz zu bisherigen Videoanalysetools, die entweder nur manuelle Annotationen erlauben (z. B. *ANVIL*[16] von Kipp 2014; *Cinemetrics*[17] von Tsivian 2009; *ELAN*[18] von Wittenburg et al. 2006) oder nur wenige ausgewählte Methoden zur automatischen Inhaltsanalyse enthalten (z. B. *Videana* von Ewerth et al. 2009; *VIAN*[19] von

15 S. https://github.com/TIBHannover/tibava, zuletzt aufgerufen am 22.06.2024.
16 S. http://www.anvil-software.de, zuletzt aufgerufen am 22.06.2024.
17 S. https://cinemetrics.uchicago.edu, zuletzt aufgerufen am 22.06.2024.
18 S. https://archive.mpi.nl/tla/elan, zuletzt aufgerufen am 22.06.2024.
19 S. https://www.vian.app, zuletzt aufgerufen am 22.06.2024.

Halter et al. 2019), stellt TIB AV-A eine umfangreiche Sammlung modernster Mustererkennungsansätze zur Verfügung, ohne dass fortgeschrittene technische Kenntnisse oder spezifische Hardwareanforderungen erforderlich sind. Nutzer*innen aus verschiedenen Disziplinen können einfach ihre eigenen Videos hochladen und haben dann Zugang zu einer Vielzahl von Analyseperspektiven. Einen Überblick über die derzeit unterstützten Methoden zur Filmanalyse gibt Tab. 1.

Tab. 1 Übersicht über aktuelle Methoden zur Bild- und Videoanalyse sowie zur Audio- und Sprachanalyse in TIB AV-A.

Bild- und Videoanalyse	Grundlegende Bildmerkmale: dominante **Farbe(n) und Helligkeit**
	Erkennung von Einstellungsgrenzen
	Schnittfrequenz (vgl. Redfern 2022a), d. h. die Häufigkeit von Einstellungswechseln
	Klassifizierung der Einstellungsgröße von Detailaufnahme, Naheinstellung, Halbtotale, Totale und großer Totale
	Klassifizierung des Ortes (z. B. Kirche, Markt, Restaurant usw.)
	Personenerkennung anhand eines Beispielbildes
	Personen-Clustering zum automatischen Auffinden der am häufigsten vorkommenden Personen/Akteure
	Erkennung von Gesichtsausdrücken (z. B. wütend, glücklich)
	Zero-Shot-Bildklassifizierung für beliebige visuelle Konzepte auf Grundlage von Textbeschreibungen (z. B. „Ein Foto aufgenommen in einem Zug", siehe Abb. 1)
	Zero-Shot-Videoklassifizierung für beliebige audiovisuelle Konzepte auf Grundlage von Textbeschreibungen (z. B. „Ein Video mit feiernden Menschen")
	Bildunterschriften zur automatischen Beschreibung von Bildern in einem Video
Audio- und Sprachanalyse	Grundlegende Audiomerkmale: **Amplitudenkurve** (Wellenform), **Lautstärke** (*Root Mean Square)* und das **Frequenzspektrum**
	Spracherkennung zur automatischen Transkription von Sprache in Videos

Neben einigen Standard-Analyseaufgaben (z. B. Farbanalyse, Erkennung von Einstellungsgrenzen) ist in TIB AV-A vor allem die Hinzufügung von Spracherkennung und Zero-Shot-Bild- und Videoklassifizierung hervorzuheben. Qualitativ hochwertige Transkripte (d. h. mit einer niedrigen Wortfehlerrate) ermöglichen eine viel bessere Analyse von Sprache mit Hilfe von Ansätzen aus der Verarbeitung natürlicher Sprache für Aufgaben wie *Topic Modeling, Named Entity Linking* usw., die in TIB AV-A zukünftig ergänzt werden. Darüber hinaus ermöglicht die Zero-Shot-Klassifikation von Bildern und Videos verschiedene nachgelagerte Aufgaben. Basierend auf einer textuellen Eingabeaufforderung können die zugrundeliegenden Bildsprachmodelle,

d. h. *CLIP* (Radford et al. 2021) und *InstructBLIP* (Dai et al. 2023) (eine Reihe von) beliebige(n) Konzepte(n) erkennen. Auf diese Weise können Anwender*innen Videos automatisch nach verschiedenen Konzepten durchsuchen, die von Objekten der realen Welt (z. B. Flaggen, Autos usw.) und Tieren über Umgebungsbedingungen (z. B. Orte, Wetter, Tageszeiten) bis hin zu viel komplexeren Konzepten reichen, z. B. Berufe von Personen (z. B. Polizist*innen, Reporter*innen), Ereignisse (z. B. Naturkatastrophen, Demonstrationen, Sportarten) usw.

Obwohl TIB AV-A eine breite Palette an modernen Methoden für die automatische Filmanalyse bietet, sind DH-Forscher*innen oft an fortgeschritteneren Mustern interessiert, die eine Kombination von Merkmalen umfassen können. Beispielsweise können Sequenzen in Filmen mit hoher *Einstellungsdichte,* plötzlichen *Lautstärkeänderungen* und *Nahaufnahmen* auf spannende Schlüsselszenen oder Aktionen in Filmen hinweisen. Die Kombination von Merkmalen kann auch Bedingungen zu bestimmten Mustern hinzufügen, um z. B. nach Aktionen zu suchen, wenn eine bestimmte Person oder ein bestimmtes Objekt sichtbar ist (siehe Abb. 1). Um solche Kombinationen zu ermöglichen, bietet TIB AV-A die Möglichkeit, Wahrscheinlichkeiten bestimmter Merkmale (z. B. Szenen, Emotionen, Einstellungsgrößen) mit logischen Operationen *(oder, und)* zu verknüpfen. Auf Grundlage der aus einem gegebenen Video extrahierten Merkmale können Benutzer interaktive Visualisierungen für die qualitative Analyse erstellen. Derzeit unterstützt TIB AV-A eine Wordcloud-Visualisierung basierend auf extrahierten Sprachtranskripten sowie Streu- und Liniendiagrammen, für die Nutzer*innen bestimmte Merkmale und Merkmalskombinationen anzeigen (und ausblenden) können (siehe Abb. 1). Darüber hinaus können Graphenvisualisierungen erstellt werden, die z. B. Personenkonstellationen und deren Vorkommen an bestimmten Orten und Plätzen zeigen.

5. Fallstudie: Analyse des Serienendes von *Game of Thrones* bezüglich narrativer Muster

Im vorangegangenen Abschnitt wurde deutlich, wie der aktuelle Stand der Forschung automatischer Analyse von Filmen eine Vielzahl von Analysemethoden unterstützt. Die Komponenten, die in TIB AV-A integriert werden, decken zwei Hauptarten ab: erstens die automatische Analyse von Filmen in Bezug auf Kategorien und Eigenschaften, die für alle Filme gelten, wie z. B. Einstellungsgrenzen, Farbbereiche, Tonspektrogramme u. ä., und zweitens die automatische Analyse von Filmen in Bezug auf Kategorien, semantische Konstrukte oder formale Merkmale, die vom Menschen ausgewählt werden. In beiden Bereichen ist zu erwarten, dass die Genauigkeit, Präzision und Vielfalt der gelieferten Ergebnisse in den kommenden Jahren erheblich zunehmen wird. Es bleiben jedoch einige Fragen offen, wie diese Fähigkeiten genutzt

werden können, um die verschiedenen Arten von Analysen zu unterstützen, die für Filme in Frage kommen. Dieser Abschnitt zeigt ein Beispiel für eine Analyse, die sich speziell auf die Aufdeckung größerer narrativer Strukturen konzentriert.

Um die Diskussion zu konkretisieren, wird die Analyse anhand der Schlussszenen der letzten Folge der finalen Staffel von *Game of Thrones* durchgeführt, die von David Benioff und D. B. Weiss für HBO entwickelt und 2019 erstmals ausgestrahlt wurde.[20] Das „Was" dieses Abschnitts ist schnell beschrieben: Die drei Hauptfiguren der handlungsrelevanten Stark-Familie der Geschichte, Jon Snow, Arya Stark und Sansa Stark, beginnen neue Abschnitte ihres Lebens. Jon Snow überschreitet die Grenze, die die Zivilisation vom eisigen Norden trennt, Arya Stark segelt nach Westen, um nach neuen Ländern zu suchen, und Sansa Stark wird zur Königin gekrönt. Damit endet die Serie. Filmisch jedoch bedient sich die Darstellung dieser Ereignisse einer Reihe von bekannten Techniken, die einen eng strukturierten Vergleich der jeweiligen Schicksale der dargestellten Personen ergeben. Aus der Sicht der Filmanalyse ist es daher wichtiger, sich mit der Frage nach dem „Wie" der Konstruktion dieses Segments zu befassen.

5.1 Workflow

Es soll nun gezeigt werden, wie die Verwendung der TIB AV-A-Plattform die Erforschung dieser Art von filmischen Mustern unterstützen kann, indem zunächst die interne Struktur des Segments aufgezeigt wird und dann kurz erörtert wird, wie diese in eine umfassendere Untersuchung der Filmform einbezogen werden kann. Es wird außerdem betont werden, wie die Arbeit mit der Ästhetik und Poetik der Filmanalyse dazu beiträgt, Prioritäten für die Implementierung von Merkmalen zu setzen, die für die schrittweise Überführung von manuellen und halbautomatischen Analysen in eine vollautomatische Analyse von Nutzen wären. Im folgenden Arbeitsablauf wird auch das ELAN-Tool für die manuelle Annotation und die Korrektur der automatischen Annotationen sowie einige benutzerdefinierte R-Skripte für die Visualisierung der Ergebnisse verwendet.

Der erste Schritt besteht darin, das Filmsegment in TIB AV-A zu laden und die standardmäßigen automatischen Verarbeitungspipelines, wie etwa für die Segmentierung und Skalierung von Einstellungen, durchzuführen. An dieser Stelle können auch Elemente, von denen bekannt ist, dass sie für das Segment von besonderer Relevanz sind, für bestimmte Kategorien verwendet werden – zum Beispiel die Suche nach Gesichtern der Hauptfiguren auf Grundlage hochgeladener Bilder oder durch die Verwendung natürlichsprachlicher Phrasen für die inhaltsbasierte Segmentierung mithilfe von zero-shot-Verfahren.

20 Die analysierte Szene ist zu sehen unter: https://www.youtube.com/watch?v=zUZvYAjaEZk, zuletzt aufgerufen am 22.06.2024.

Der zweite Schritt besteht darin, die Analyseebenen aus TIB AV-A zu exportieren und sie in eine Form zu übersetzen, die für die weitere Segmentierung und manuelle Annotation mit ELAN oder ähnlichen Tools geeignet ist. Dieser letzte Schritt wird hier lokal mit speziellen Verarbeitungsskripten durchgeführt. Dies ermöglicht die Korrektur von Fehlern bei der automatischen Verarbeitung sowie das Hinzufügen weiterer filmischer Merkmale, die von TIB AV-A noch nicht automatisch bereitgestellt werden. Relevante Beispiele hierfür sind im vorliegenden Fall Kamerabewegungen, da das Segment in hohem Maße auf die Kohäsion der Kamerabewegungen über Teilsequenzen hinweg angewiesen ist. Das allgemeine Analyseschema folgt dann dem in Bateman & Schmidt (2012) dargelegten Schema, bei dem Aufnahmen räumlich-zeitlichen Regionen zugeordnet werden. Die menschliche visuelle Wahrnehmung entscheidet im Allgemeinen sehr schnell und genau, ob sie einen bestimmten Ort schon einmal gesehen hat, und diese Art von Kontinuität ist auch aus psychologischen Studien als grundlegende Einheit für das erweiterte Diskursverständnis bekannt (Zacks 2010; Loschky et al. 2020). In ELAN werden folglich Annotationsebenen definiert, denen Aufnahmen zugeordnet werden können. Dies ist ein Bereich, in dem eine immer genauere Szenenerkennung in Kombination mit visuellen Ähnlichkeitsmaßen in naher Zukunft wesentliche Verbesserungen zur Unterstützung der automatischen oder halbautomatischen Analyse erwarten lässt. Auf diese Weise wird sichtbar, wie Fragen, die sich direkt aus den Bedürfnissen der poetologischen und ästhetischen Analyse von Filmen ergeben, nach und nach von den sich entwickelnden computergestützten Werkzeugen übernommen und unterstützt werden können. Fehlende Merkmale können zunächst manuell hinzugefügt und dann computergestützt unterstützt werden, wenn sie verfügbar werden.

Der dritte und für unsere Zwecke letzte Schritt besteht darin, die ELAN-Analyse weiter zu exportieren, um wiederkehrende filmische Muster gezielt zu untersuchen. Hierfür werden eigens erstellte R-Skripte verwendet, die lokal laufen und die ELAN-Annotationen direkt in Visualisierungen der filmischen Struktur umwandeln, welche je nach Wunsch mit Ergebnissen der automatischen und manuellen Analyse überlagert werden. Während viele klassische formale Schnittmerkmale in R mittlerweile auf recht ausgefeilte Weise auf ihre statistischen Eigenschaften hin untersucht werden können (vgl. Redfern 2022b), geht es hier eher um die Ableitung übergeordneter organisatorischer Eigenschaften, die oft direkter mit Interpretationen korrespondieren. Die hier verwendeten Visualisierungen sind in Bateman & Schmidt (2012) definiert und lehnen sich lose an die musikalische Notation an, indem sie aufeinanderfolgende Einstellungen horizontal anordnen, sodass weitere strukturelle Beziehungen, Eigenschaften und Gruppierungen frei hinzugefügt werden können. Kurz gesagt, es wird hier versucht, funktional relevante Sequenzen von Kombinationen filmischer Merkmale zu identifizieren, die über allgemeine Statistiken von Übergängen, Koinzidenzen u. ä. hinausgehen können (Bateman 2014).

5.2 Analyse

Die Grundstruktur des Beispielabschnitts ist in Abb. 2 dargestellt. Sie zeigt die horizontal nummerierten Einstellungen der Szene in der unteren Reihe und kurze funktionale Beschreibungen dieser Einstellungen in der oberen Reihe, um die Orientierung zu erleichtern. Wann immer eine Einstellung bestimmten funktionalen Gruppierungen entspricht, wird sie weiter in „Untereinstellungen" unterteilt – wie z. B. in Einstellung 15, die weiter in ein Segment unterteilt ist, das eine gehende Figur verfolgt (15.1), gefolgt von einem stationären Fokus auf diese Figur (15.2).

Anhand dieser Visualisierung lässt sich die im Wesentlichen dreizeilige Entwicklung der Sequenz gut erkennen, in der die aufeinanderfolgenden Einstellungen häufig die verschiedenen Schauplätze der drei Hauptfiguren (vertikal angeordnet) abdecken. Diese Struktur wird in Bateman & Schmidt (2012, 222–226) formal als dreigeteilte polyräumliche Abwechslung definiert und drückt i. d. R. Kontrast und Vergleich aus. Jede Einstellung wird hier auch mit ihrer Einstellungsgröße gekennzeichnet, die von engen Nah- oder Detailaufnahmen (orig. „tight close-up shots", TS) bis zu extra weiten Aufnahmen (orig. „extra long shots", ELS) reicht. Die Sequenz beginnt also mit drei Nahaufnahmen, die nacheinander die Schauplätze von Jon Snow (JS), Arya Stark (AS) und Sansa Stark (SS) durchlaufen; die nächsten drei Einstellungen, ebenfalls Nahaufnahmen, wiederholen diese Übergänge in umgekehrter Reihenfolge, und so weiter.

Filmisch ist es interessant, genauer zu untersuchen, wie die Konstruktion des Segments trotz dieser schnellen Übergänge zwischen den Szenen die Kohärenz aufrechterhält. Um dies zu untersuchen, wird die Visualisierung sukzessive um weitere Informationsschichten aus den Annotationen erweitert. Abb. 3 zeigt z. B. die Visualisierung mit eingeblendeten Annotationsebenen zu verschiedenen Arten der Kameranutzung, sowohl als Beschriftungen als auch als farbige Gruppierungen über den betroffenen Einstellungen. Die Einstellungen, die in dieser Abbildung nach der *Kamerarichtung* klassifiziert sind, zeigen gut, wie die Richtung die Kohäsion über die verschiedenen Schauplätze hinweg aufrechterhält. Einstellungen 18–22 bspw. halten die Richtung nach rechts, während in den Einstellungen 23–25 die Kamera aus dem Szenenraum herausbewegt wird. Aufeinanderfolgende Einstellungen innerhalb desselben Schauplatzes (Einstellungen 28–30, 38–40, 42–46 in Abb. 2) erscheinen im Gegensatz dazu ohne erkennbare Kamerarichtung. Und entscheidend ist, dass keines dieser filmtechnischen Merkmale allein die Bedeutung der Konstruktion trägt; erst in ihrer strukturellen Zusammensetzung entsteht eine verlässlich interpretierbare Form.

Die farbigen Balken am oberen Rand des Diagramms zeigen auch, wie andere Dimensionen der Kameranutzung, hier Bildausschnitt und Bewegung, ebenfalls dazu dienen, Einstellungen zu gruppieren, wiederum häufig über die drei Schauplätze hinweg. Die unteren Symbollinien des Diagramms zeigen an, welche Arten von Cadrage und Bewegung am feinsten ausgeprägt sind; Einstellungen 19–22 zeigen z. B. alle eine konstante Rahmung (S: „Seitenblick"), die zum Zusammenhalt der Sequenz beiträgt.

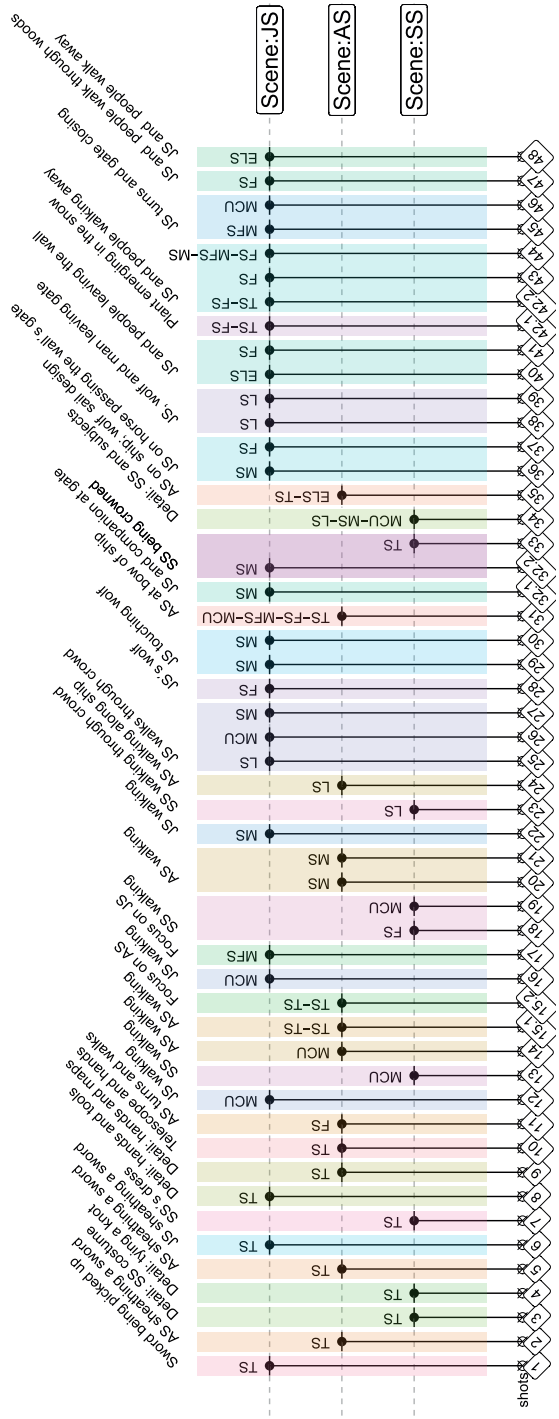

Abb. 2 Grundlegende Visualisierung der annotierten Filmstruktur des *Game of Thrones*-Segments; die Einstellungen verlaufen horizontal, unten sind sie nummeriert; die Einfärbung zeigt die Gruppierung des semantischen Inhalts an (alle Diagramme wurden mit dem R-Paket *ggplot* erstellt, Wickham 2016). Die Abkürzungen der Shot-Skalen basieren auf Standard-Shot-Größen in zunehmender Entfernung: *Tight* oder *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).

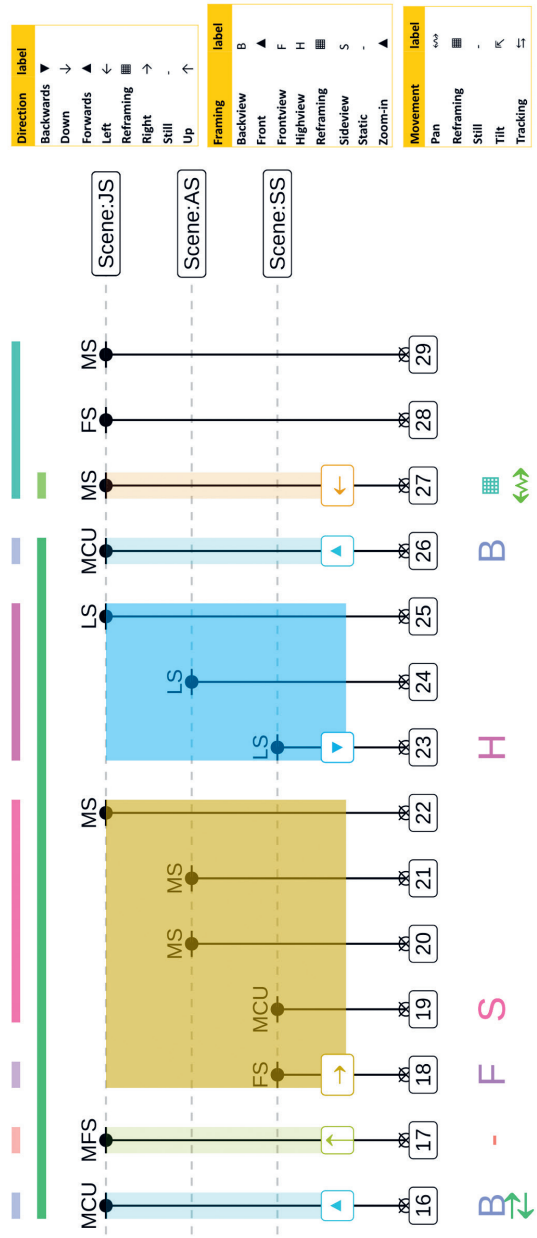

Abb. 3 Visualisierung der kommentierten Filmstruktur des *Game of Thrones*-Segments (Einstellungen 15 – 29), ergänzt durch Informationen zur Kameranutzung. In dieser Visualisierung wurde die Kamerarichtung priorisiert und die Gruppierung mit größeren farbigen Blöcken dargestellt. Die Balken oben zeigen die Gruppierung, die durch Kadrierung und Bewegung erzwungen wird; die Symbole unten zeigen, um welche Art von Kameranutzung es sich jeweils handelt.

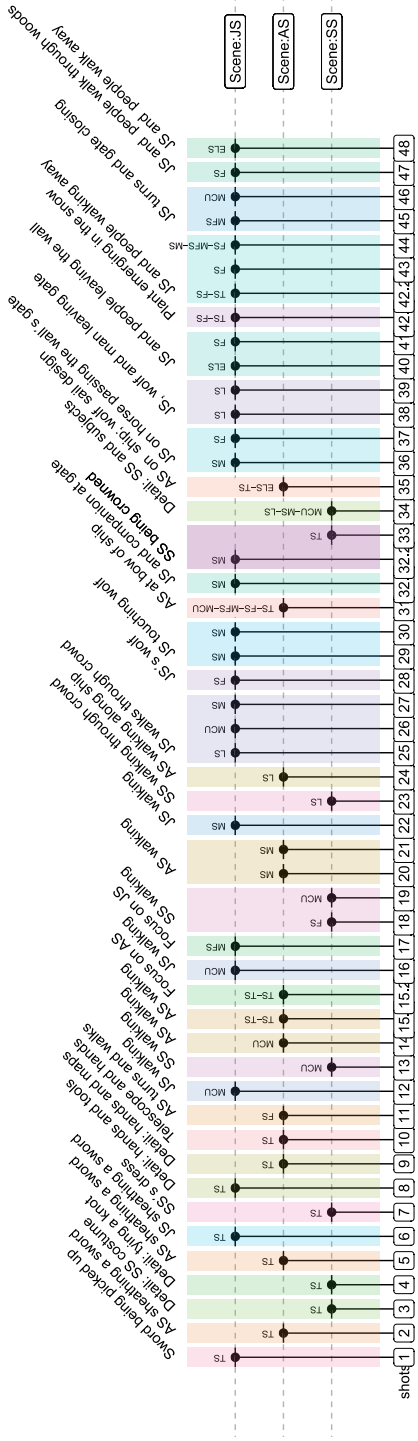

Abb. 4 Visualisierung der annotierten Filmstruktur des *Game of Thrones*-Segments mit den Ergebnissen der automatischen Gesichtserkennung und der CLIP-basierten semantischen Konzepterkennung.

Jede dieser filmischen Eigenschaften kann zur visuellen Hervorhebung in den Visualisierungen ausgewählt werden, sodass die verschiedenen Arten von Strukturen sichtbar gemacht werden können. Relevante Beispiele wären hier die kontinuierliche diegetische Tonanpassung über die Szenen hinweg (z. B. Schritte) sowie die übergreifende nicht-diegetische Musik, die das Segment gliedert, wobei das *Stark-Familienmotiv* in den Einstellungen 1–17 durchläuft und sich allmählich mit dem *Game of Thrones*-Thema in den Einstellungen 18–30 vermischt, welches anschließend in den verbleibenden Einstellungen 31–48 dominiert.

Es ist auch möglich, andere von TIB AV-A erhaltene automatische Annotationsergebnisse zu überlagern. Abb. 4 zeigt zum Beispiel, wo TIB AV-A eines der drei Hauptgesichter der mit hoher Sicherheit auftretenden Charaktere (klassifiziert durch visuelle Ähnlichkeit) und das Auftreten von Jon Snows Wolf (klassifiziert durch semantische zero-shot-Klassifizierung mit CLIP, wie oben beschrieben) aufgezeichnet hat. Interessant für die filmische Konstruktion der Sequenz ist dabei, dass die Identifikation der Protagonisten erst recht spät erfolgt: Erst ab Einstellung 15 werden Gesichter gezeigt, während sich die Geschichte auf ihr Finale zubewegt. Auch die Platzierung des Wolfs in der Mitte der Szenen mit Jon Snow ist recht genau.

Während die bisher gezeigten Ansichten eine weitere Erforschung des Aufbaus dieser Sequenz unterstützen, wird es zukünftig von Vorteil sein, strukturelle Muster auf Grundlage der aufgedeckten Strukturen zu definieren, die dann wiederum in die automatischen Suchfunktionen von TIB AV-A und anderen Werkzeugen zurückgeführt werden können. Dies erfordert die Definition von Mustern als Suchanfragen. Im vorliegenden Fall würde man z. B. nach sich wiederholenden Sequenzen von Einstellungen suchen, die jeweils von einem anderen Ort aus aufgenommen wurden, aber dennoch eine Reihe von identischen formalen technischen Merkmalen aufweisen, wie z. B. Kamerabewegung, Richtung usw. Die Ausweitung solcher Musterabfragen auf alle möglichen automatisch ermittelten Merkmale verspricht, den *state of the art* für die computergestützte Filmanalyse im großen Maßstab drastisch zu verändern und den Kontakt zu eher hermeneutisch orientierten Forschungsaufgaben wiederherzustellen.

6. Fazit und künftige Herausforderungen

Bislang wurde bei der Analyse von Film und Video in den DH schon viel erreicht. Mit dem Durchbruch von Deep-Learning-Ansätzen in den letzten Jahren ist auch eine breite Palette von Methoden entstanden, die die automatische Extraktion verschiedener multimodaler Merkmale erleichtern. Ein erheblicher Anstieg der quantitativen Datenverfügbarkeit macht die Entwicklung eines entsprechenden analytischen Rahmenwerks erforderlich. Wir sind der Meinung, dass ein solcher Rahmen sowohl auf empirischen Standards als auch auf theoretischen Grundlagen wie der multimodalen

Theorie und der Semiotik beruhen sollte. Er würde auch von der Integration von Konzepten aus gängigen Taxonomien profitieren, die von Forschenden in den DH verwendet werden, wie z. B. die AdA-Filmontologie[21], mit der erste vielversprechende Versuche mit TIB AV-A durchgeführt worden sind. Eine geeignete Integration, die auch die hierarchische Natur solcher Ontologien erfasst, muss jedoch noch entwickelt werden.

Während die empirische Analyse als Eckpfeiler der computergestützten Filmanalyse angesehen werden kann, gibt es ein zentrales Argument für die Nutzung von Explorationswerkzeugen wie TIB AV-A. Bisher konzentrieren sich die meisten existierenden Werkzeuge auf die Erkundung und Visualisierung einzelner Videos. Um das Konzept des Distant Viewing (Arnold & Tilton 2019) über mehrere Videos hinweg zu realisieren, müssen wir Methoden zur gleichzeitigen Visualisierungsanalyse entwickeln. Dies ist aufgrund der dynamischen Natur von Videoinhalten eine schwierige Aufgabe. Erste Schritte in diese Richtung wurden bereits durch Cultural Analytics (Manovich 2020) und Visual Movie Analytics (Kurzhals et al. 2016) unternommen. Da der Trend in den textuellen DH jedoch in Richtung Scalable Viewing geht (Weitin 2017), d. h. eines hybriden Ansatzes, der es Forschenden ermöglicht, zwischen nah und fern fließend zu wechseln, ist dieses Konzept auch für die Analyse von Videomaterial vielversprechend. Erste Beispiele für Scalable Viewing finden sich in Ansätzen zur Visualisierung von Nachrichtenvideos (Liebl & Burghardt 2023; Ruth et al. 2023) sowie in allgemeineren Tools wie *PixPlot*[22] oder dem *Collection Space Navigator* (Ohm et al. 2023).

Literaturverzeichnis

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* (S. 23716–23736). New Orleans, Louisiana: Neural Information Processing Systems. https://doi.org/10.48550/arXiv.2204.14198 [zuletzt aufgerufen am 22.06.2024].

Arnold, T., & Tilton, L. (2019). Distant viewing. Analyzing large visual corpora, *Digital Scholarship in the Humanities,* 34(1), 3–16. https://doi.org/10.1093/llc/fqz013 [zuletzt aufgerufen am 22.06.2024].

21 S. https://projectada.github.io/ontology, zuletzt aufgerufen am 22.06.2024.

22 PixPlot, vom Yale DH Lab: https://github.com/YaleDHLab/pix-plot, zuletzt aufgerufen am 22.06.2024.

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

Dies. (2022). Analyzing Audio/Visual Data in the Digital Humanities. In J. O'Sullivan (Hrsg.), *The Bloomsbury Handbook to the Digital Humanities* (S. 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* (S. 12449–12460). Online: Neural Information Processing Systems. https://doi.org/10.48550/arXiv.2006.11477 [zuletzt aufgerufen am 22.06.2024].

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 (Hrsg.), *Visual Communication* (S. 301–330). Berlin/Boston: De Gruyter Mouton.

Ders. (2022). Growing theory for practice. Empirical multimodality beyond the case study, *Multimodal Communication,* 11(1), 63–74. https://doi.org/10.1515/mc-2021-0006 [zuletzt aufgerufen am 22.06.2024].

Ders., & 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. https://doi.org/10.13140/RG.2.2.12763.72486 [zuletzt aufgerufen am 22.06.2024].

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

Dies., & 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* (S. 3111–3115). Brno: International Speech Communication Association. https://doi.org/10.21437/Interspeech.2021-560 [zuletzt aufgerufen am 22.06.2024].

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 [zuletzt aufgerufen am 22.06.2024].

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 (Hrsg.), *DHd 2016. Modellierung – Vernetzung – Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma. Konferenzabstracts.* 2. überarbeitete und erweiterte Auflage (S. 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 [zuletzt aufgerufen am 22.06.2024].

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 [zuletzt aufgerufen am 22.06.2024].

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.

Dies., 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.* https://doi.org/10.48550/arXiv.2305.06500 [zuletzt aufgerufen am 22.06.2024].

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 (Hrsg.). *Proceedings of the European Conference on Computer Vision 2020* (S. 741–757). Cham: Springer [= Lecture Notes in Computer Science, 12356]. https://doi.org/10.1007/978-3-030-58621-8_43 [zuletzt aufgerufen am 22.06.2024].

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* (S. 4171–4186). Minneapolis: Association for Computational Linguistics. https://doi.org/10.18653/V1/N19-1423 [zuletzt aufgerufen am 22.06.2024].

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 (Hrsg.), *Digital Tools in Media Studies. Analysis and Research. An Overview* (S. 101–116). Bielefeld: Transcript Verlag.

Flückiger, B. (2011). Die Vermessung ästhetischer Erscheinungen, *Zeitschrift für Medienwissenschaft,* 3(2), 44–60. https://doi.org/10.1524/zfmw.2011.0022 [zuletzt aufgerufen am 22.06.2024].

Dies., & 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 [zuletzt aufgerufen am 22.06.2024].

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.* https://doi.org/https://doi.org/10.48550/arXiv.2203.05794 [zuletzt aufgerufen am 22.06.2024].

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. https://doi.org/https://doi.org/10.1111/cgf.13676 [zuletzt aufgerufen am 22.06.2024].

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* (S. 770–778). Las Vegas: Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/CVPR.2016.90 [zuletzt aufgerufen am 22.06.2024].

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* (S. 2496–2500). Bordeaux: Institute of Electrical and Electronics Engineers. https://doi.org/10.48550/arXiv.2202.12555 [zuletzt aufgerufen am 22.06.2024].

Howanitz, G. (2015). Distant Waching. Ein quantitativer Zugang zu YouTube-Videos. In *DHd 2015. Von Daten zu Erkenntnissen. Book of Abstracts* (S. 1–6 [33–38]). Graz: Verband Digital Humanities im deutschsprachigen Raum. URL: https://gams.uni-graz.at/o:dhd2015.abstracts-gesamt [zuletzt aufgerufen am 22.06.2024].

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 [zuletzt aufgerufen am 22.06.2024].

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

(Hrsg.), *Proceedings of the European Conference on Computer Vision 2020* (S. 709–727). arXiv: Institute of Electrical and Electronics Engineers. https://doi.org/https://doi.org/10.48550/arXiv.2007.10937 [zuletzt aufgerufen am 22.06.2024].

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

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

Korte, H. (2004). *Einführung in die systematische Filmanalyse. Ein Arbeitsbuch.* 3. Aufl. 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* (S. 3791–3794). arXiv: Association for Computing Machinery. https://doi.org/10.48550/arXiv.2108.06543 [zuletzt aufgerufen am 22.06.2024].

Kurzhals, K., John, M., Heimerl, F., Kuznecov, P., & Weiskopf, D. (2016). Visual Movie Analytics, *IEEE Transactions on Multimedia,* 18(11), 2149–2160. https://doi.org/10.1109/TMM.2016.2614184 [zuletzt aufgerufen am 22.06.2024].

Liebl, Ch., & Burghardt, M. (2023). Zoetrope. Interactive Feature Exploration in News Videos. In W. Scholger, G. Vogeler, T. Tasovac, A. Baillot, & P. Helling (Hrsg.), *Digital Humanities 2023. Collaboration as Opportunity* (S. 432–434). Graz: Alliance of Digital Humanities Organisations. https://doi.org/10.5281/zenodo.7961822 [zuletzt aufgerufen am 22.06.2024].

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. https://doi.org/10.48550/arXiv.1901.08928 [zuletzt aufgerufen am 22.06.2024].

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* (S. 15980–15989), New Orleans, LA: Institute of Electrical and Electronics Engineers. https://doi.org/10.48550/arXiv.2204.02509 [zuletzt aufgerufen am 22.06.2024].

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. https://doi.org/10.1371/journal.pone.0142474 [zuletzt aufgerufen am 22.06.2024].

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

Ders. (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. (Hrsg.), *Computer Vision. ECCV 2018* (S. 575–592). Springer, Cham [= *Lecture Notes in Computer Science,* 11216]. https://doi.org/10.1007/978-3-030-01258-8_35 [zuletzt aufgerufen am 22.06.2024].

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* (S. 2928–2938). Waikoloa: Institute of Electrical and Electronics Engineers. https://doi.org/10.48550/arXiv.2011.04714 [zuletzt aufgerufen am 22.06.2024].

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. (Hrsg.), *Proceedings of the European Conference on Computer Vision 2022* (S. 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.* https://doi.org/10.48550/arXiv.2305.06809 [zuletzt aufgerufen am 22.06.2024].

Prinz, S. (2007). *Movies and meaning. An introduction to film.* 4. Aufl. 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 [zuletzt aufgerufen am 22.06.2024].

R-Kernteam. (2016). R. Eine Sprache und Umgebung für statistische Berechnungen [Computersoftware]. *R Foundation for Statistical Computing.* URL: https://www.R-project.org/ [zuletzt aufgerufen am 22.06.2024].

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 (Hrsg.), *Proceedings of the International Conference on Machine Learning* (S. 8748–8763). arXiv. [= *Proceedings of Machine Learning Research,* 139]. https://doi.org/10.48550/arXiv.2103.00020 [zuletzt aufgerufen am 22.06.2024].

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 (Hrsg.), *Proceedings of the International Conference on Machine Learning 2023* (S. 28492–28518). Honolulu: International Machine Learning Society [= *Proceedings of Machine Learning Research,* 202]. https://doi.org/10.48550/arXiv.2212.04356 [zuletzt aufgerufen am 22.06.2024].

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.* https://doi.org/10.48550/arXiv.2106.04624 [zuletzt aufgerufen am 22.06.2024].

Redfern, N. (2022a). Analysing Motion Picture Cutting Rates, *Wide Screen,* 9(1). 1–29. URL: https://widescreenjournal.org/vol-9-no-1-2022-title [zuletzt aufgerufen am 22.06.2024].

Ders. (2022b). *Computational Film Analysis with R.* Version 0.9.004. Zenodo. https://doi.org/10.5281/ZENODO.7074521 [zuletzt aufgerufen am 22.06.2024].

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 (Hrsg.), *Proceedings of the Computational Humanities Research Conference 2023* (S. 167–177). Paris: Computational Humanities Research. [= *CEUR Workshop Proceedings,* 3558] URL: https://ceur-ws.org/Vol-3558 [zuletzt aufgerufen am 22.06.2024].

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. https://doi.org/10.2307/1211438 [zuletzt aufgerufen am 22.06.2024].

Ders. (2006). *Moving Into Pictures. More on Film History, Style, and Analysis.* London: Starword Publishing. URL: http://www.starword.com/MovPicFin.pdf [zuletzt aufgerufen am 22.06.2024].

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* [S. 1–4]. Istanbul: IEEE Xplore. https://doi.org/10.1109/ICEET53442.2021.9659697 [zuletzt aufgerufen am 22.06.2024].

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.* https://doi.org/10.48550/arXiv.2008.04838 [zuletzt aufgerufen am 22.06.2024].

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 (Hrsg.), *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval 2023* (S. 3195–3199). New York: Association for Computing Machinery Special Interest Group on Information Retrieval. https://doi.org/10.1145/3539618.3591820 [zuletzt aufgerufen am 22.06.2024].

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* (S. 750–760). Waikoloa: Institute of Electrical and Electronics Engineers. https://doi.org/10.48550/arXiv.2104.14995 [zuletzt aufgerufen am 22.06.2024].

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 (Hrsg.), *DHd 2023. Open Humanities Open Culture.* Trier/Luxemburg: Verband Digital Humanities im deutschsprachigen Raum. https://doi.org/10.5281/zenodo.7715277 [zuletzt aufgerufen am 22.06.2024].

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

Vonderau, P. (2017). Quantitative Werkzeuge. In Hagener, M., & Pantenburg, V. (Hrsg.), Handbuch Filmanalyse. Wiesbaden: Springer VS [= *Springer Reference Geisteswissenschaften*]. https://doi.org/10.1007/978-3-658-13352-8_28-1 [zuletzt aufgerufen am 22.06.2024].

Walkowski, N.-O., & Pause, J. (2018). Everything is Illuminated. Zur numerischen Analyse von Farbigkeit in Filmen, *Zeitschrift für digitale Geisteswissenschaften,* o. S. Wolffenbüttel: Herzog August Bibliothek. https://doi.org/10.17175/2018_003 [zuletzt aufgerufen am 22.06.2024].

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. https://doi.org/10.1093/llc/fqy085 [zuletzt aufgerufen am 22.06.2024].

Wickham, H. (2016). *ggplot2. Elegant Graphics for Data Analysis.* Berlin/Heidelberg: Springer. https://doi.org/10.1007/978-3-319-24277-4 [zuletzt aufgerufen am 22.06.2024].

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* (S. 1556–1559). Genoa: ELRA Language Resources Association. URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/153_pdf.pdf [zuletzt aufgerufen am 22.06.2024].

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* (S. 10669–10680). Vancouver: Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/CVPR52729.2023.01028 [zuletzt aufgerufen am 22.06.2024].

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. https://doi.org/10.1145/3241057 [zuletzt aufgerufen am 22.06.2024].

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* (S. 6397–6407). arXiv: Association for Computational Lingutistics. https://doi.org/10.48550/arXiv.1911.03814 [zuletzt aufgerufen am 22.06.2024].

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* (S. 1–5). Rhodes Island: IEEE. https://doi.org/10.1109/ICASSP49357.2023.10095969 [zuletzt aufgerufen am 22.06.2024].

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* (S. 114–123). Seattle: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-demo.12 [zuletzt aufgerufen am 22.06.2024].

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* (S. 543–553). Singapur: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-demo.49 [zuletzt aufgerufen am 22.06.2024].

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. https://doi.org/10.1109/TPAMI.2017.2723009 [zuletzt aufgerufen am 22.06.2024].

Bildnachweise

Bei den Abbildungen 1–4 handelt es sich um selbst erstellte Screenshots aus der Arbeit der Autoren mit TIB AV-A (Abb. 1) und dem *R*-Paket *ggplot* (Abb. 2–4). Sie alle werden hier erstveröffentlicht.