A06 Forschungsüberblick: Distant Viewing, Computer Vision und DH

Stand: 05. Mai 2026

In diesem Dokument werden Rechercheergebnisse zu computergestützter Verarbeitung und Analyse filmischer und bildlicher Daten für das A06 Projekt festgehalten und kontinuerlich ergänzt und erweitert.

Definitionen und Theorie

In einem ersten Schritt werden zentrale Begriffe und Themen im Kontext computergestützter Bild- und Filmverarbeitung vorgestellt, wobei der Fokus auf den digitalen Geisteswissenschaften (DH) liegt.

Distant Viewing

Distant Viewing bezeichnet „the application of computer vision methods to the computational analysis of digital images“ (Arnold und Tilton 2023, S. 11). Distant Viewing ist eng mit der Bildsemiotik und Informationswissenschaft verbunden, da es um die Frage geht, wie in einem Bild Bedeutung kodiert und anschließend wieder dekodiert/verstanden werden kann. Laut Arnold und Tilton (2023, S. 14ff) gibt es zentrale Unteschiede zwischen Bildern und Schrift/Sprache, was folglich auch Unterschiede zwischen Distant Reading (Underwood 2017) und Distant Viewing mit sich bringt. Inwieweit dies zutreffend ist, soll an dieser Stelle nicht weiter diskutiert werden. Laut Arnold und Tilton sind Bilder anders als Schrift über das „sharing [of] similar qualities“ miteinander verbunden (Arnold und Tilton 2023, S. 14), was für Schrift und Sprache als Symbolsystem so nicht gilt. Allerdings werden Bilder basierend auf kultureller Prägung und individueller Erfahrung unterschiedlich dekodiert und damit interpretiert. Gerade dieser Aspekt der kulturell geprägten Dekodierung von Bildern stellt eine Herausforderung bei der Entwicklung von Distant-Viewing-Verfahren dar (Arnold und Tilton 2023, S. 16).

Beim Distant Viewing ist es notwendig, aus den Bilddaten (Pixel) in einem ersten Schritt Informationen zu dekodieren (Annotationen, Labels), die dann erst in einem zweiten Schritt weiter aggregiert und zusammenhängend analysiert werden können, um das Bild zu verstehen, was entweder ebenfalls durch einen Computer oder aber einen Menschen geschehen kann.

> *To be more concrete, the computational analysis of a collection of digital images requires as a first step that the messages encoded within the images are interpreted through the construction of structured labels.* (Arnold und Tilton 2023, S. 23)

Im Zentrum von Distant Viewing steht folglich die Erstellung und Analyse von mit der Hilfe von Computer Vision ausgezeichneter Bilddaten (Annotationen)[1], wobei Computer Vision die Decodierung von viel größeren Bildmengen erlaubt, als dies für einen Menschen möglich wäre, was wiederum die Basis für das anschließende Verstehen verändert. Ein kritischer Punkt in diesem Kontext fällt der Frage zu, wie objektiv die Dekodierung von Bildern durch computergestützte Verfahren geschieht bzw. durch welche menschlichen/kulturellen Faktoren diese letztlich geprägt ist (Arnold und Tilton 2023, S. 28).

Computer Vision

> *The field of computer vision focuses on how computers automatically produce ways of understanding digital images.* (Arnold und Tilton 2023, S. 26)

Das Feld der Computer Vision umfasst verschiedene Disziplinen (Informatik, Psychologie usw.) und Ansätze, die zur computergestützten Annotation (dem *Dekodieren*, siehe Abschnitt über Distant Viewing) sowie Analyse von Bild- und Filmdaten verwendet werden. Die angewendeten Verfahren und Programme stammen häufig aus Anwendungsbereichen, die in keinem direkten Zusammenhang mit universitärer bzw. geisteswissenschaftlicher Forschung stehen, sondern aus der Industrie oder Überwachungssystemen stammen. Darunter fallen:

1. Bildsegmentierung (Annotation zusammenhängender Bildbereiche, bspw. eines Objekts).
2. Erkennung und Tracking von Objekten (bspw. von Personen über mehrere Bilder hinweg).
3. Bildklassifizierung in bestimmte Kategorien.
4. Bildgenerierung.

Alle genannten Verfahen sind häufig an eine Nachahmung menschlichen Sehens angelehnt (*mimic the human visual system*). Abseits des *Dekodierens* bildlicher Informationen durch einen Computer stellt insbesondere der zweite Schritt des kulturspezifischen Verstehens von Annotationen (bspw. von annotierten Emotionen) ein komplexes Thema dar und es muss von Fall zu Fall entschieden werden, wie sinnvoll Computer Vision insbesondere für den Verstehensschritt angewendet werden kann. Computer Vision wird auch in multimodalen Kontexten, bestehend aus Texten und Bilddaten, eingesetzt sowie zur Augmentierung visueller Medien, beispielsweise durch die Generierung von Bildlabels.

Was die konrketen Verfahren betrifft, so sind - wie in anderen Gebieten auch - sowohl klassische regelbasierte Ansätze (wie Mustererkennung) als auch maschinelle Lernverfahren vorzufinden (bspw. Convolutional Neural Networks [CNN] und Transformer-basierte Ansätze wie Vision Transformer), wobei letztere in den vergangenen Jahren immer populärer geworden sind.

1 „The process of creating and interpreting digital images through annotations generated by computer vision is at the center of distant viewing.“ (Arnold und Tilton 2023, S. 24)

Forschungsstand und Methode

Arnold und Tilton präsentieren in ihrem Buch zu „Distand Viewing“ (2023) einen prototypischen Workflow zur Integration von Computer Vision in die Bild- bzw. Filmanalyse, der eng an Workflows aus dem Bereich Data Science angelehnt ist (S. 34ff.). Nicht thematisiert bzw. bereits vorausgesetzt wird im Folgenden der Schritt der Datensammlung und -speicherung sowie das Design einer Forschungsfrage.

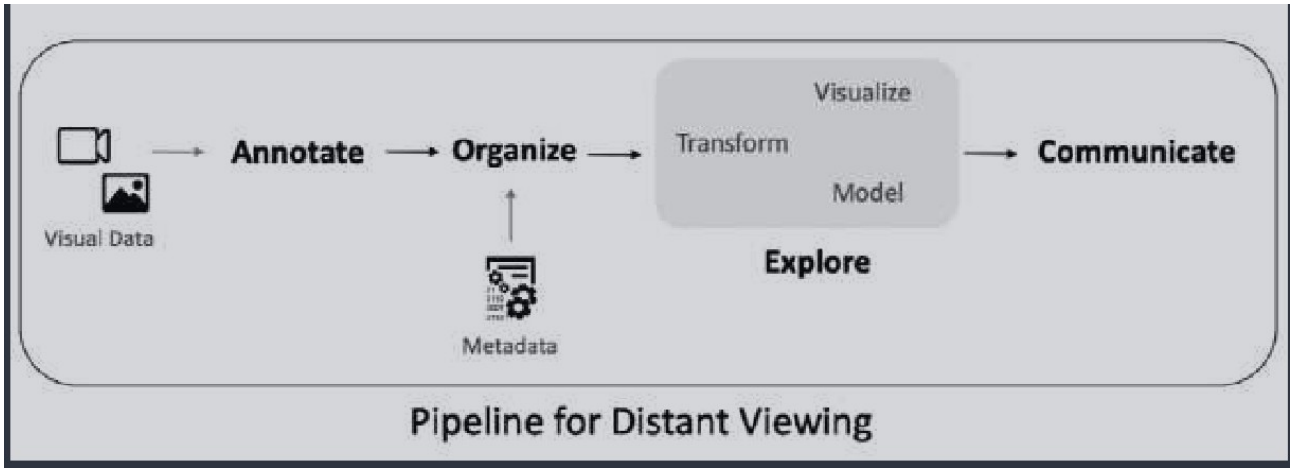

Schaubild 1: Aus: Arnold und Tilton 2023, S. 35.

1. Der erste Schritt der **Annotation** bezieht sich auf die notwendige Dekodierung von Bildinformationen (siehe Kapitel über Distant Viewing), bspw. durch die Auszeichnung von Bildausschnitten, Objekterkennung o.ä. Die Annotationen bilden das Fundament für alle weiteren Analysen. Die Auswahl der Annotationen hängt dabei stark von der Forschungsfrage ab und kann sehr verschiedene Kategorien umfassen, darunter einzelne numerische Werte (Anzahl an Personen im Bild, Helligkeit des Bilds etc.), Objekterkennungen, Bildbeschreibungen, Segmentdaten von Bildbereichen und vieles mehr. In der Regel sollte für diesen Schritt der automatisierten Annotation auf bereits existierenden Lösungen/Modellen aufgebaut werden, weil das Training/Erstellung eines eigenen automatisierten Annotationsmodells sehr langwierig bis – nicht zuletzt abhängig von der Datenmenge – unmöglich sein kann (siehe dazu die Sektion Tools).
2. Der anschließende Schritt der **Organisation** umfasst die Aggregation der Annotationen und deren Zusammenführung mit Metadaten. Hierbei können zusätzliche Bildinformationen wie Provenienz, Größe usw. mit den Annotationsdaten gemergt werden. Außerdem können auf Basis von Ähnlichkeitsvergleichen oder überschneidenden Annotationen Gruppen von Bildern mit ähnlichen Eigenschaften im Gesamtcorpus identifiziert werden, was wiederum zur Vervollständigung der Annotationen aller Bilder in den jeweiligen Gruppen genutzt werden kann.
3. Der **Explorationsschritt** umfasst einen iterativen Loop aus Datentransformationen, -analysen (bzw. Modellbildungen) und -visualisierungen, die im Einzelnen stark von der

konkreten Forschungsfrage abhängen und bspw. das Training eines ML-Modells zur Erkennung von räumlichen Bewegungen von Schauspielenden im Raum und dessen Auswertung im Kontext einer geisteswissenschaftlichen Forschungsfrage beinhalten könnten. Der Explorationsschritt enthält außerdem eine *explorative Datenanalyse* (EDA), die u.a. dazu genutzt werden kann, einen statistischen Überblick über die Daten zu erhalten sowie auf etwaige Datenfehler oder besonders auffällige Annotationen aufmerksam zu werden („Wieso finden sich in den Bildern 51-55 im Kontext eines Videos über den Vatikan plötzlich viele Annotationen von XYZ, die sonst nie vorkommen?“), was wiederum für die übergreifende Fragestellung relevant ist.

4. Oft vernachlässig, aber insbesondere für Arbeiten im interdisziplinären Rahmen von zentraler Bedeutung ist schließlich der letzte Schritt der **Kommunikation der Ergebnisse**, der die häufig komplexen und schon aufgrund ihrer Menge schwer fassbaren Ergebnisse computergestützter Analysen für ein breiteres Publikum aufbereitet und vermittelt, wobei Visualisierungen eine zentrale Rolle spielen.

In der Anfangsphase der automatisierten Analyse von Filmdaten, auch als Cinemetrics bekannt[2], lag ein Schwerpunkt auf der Annotation und Analyse von zusammenhängenden Sequenzen und deren Auswertung, darunter Szenenlänge (Butler 2014; Salt 1974), Farbgebung oder sprachliche Aspekte (Burghardt u. a. 2016). Jüngere Beiträge seit ca. 2015 benutzen verstärkt Methoden aus dem Bereich des maschinellen Lernens, um mithilfe von Verfahren wie Gesichts- oder Charaktererkennung vielfältige Forschungsfragen u.a. aus dem Bereich Gender/Race zu bearbeiten (Arnold u. a. 2019; Bamman u. a. 2024). Häufig geht es dabei darum, bestimmte Personen (Schauspieler) zu erkennen und ihre Screentime zu berechnen, beispielsweise um daraus Rückschlüsse auf die Präsenz von männlichen/weiblichen Schauspieler:innen über die Zeit ziehen (sogenannte *shot semantics*, siehe Arnold u. a. 2019, S. 5).

Tools

- **Distant Viewing Toolkit**: Das Distant Viewing Toolkit ist ein Python-Paket, das die computergestützte Analyse von Bild- und Filmdaten erleichtert. Die aktuellste Version des Pakets konzentriert sich darauf, eine minimale Menge an Funktionen bereitzustellen, um solche Arbeiten durchzuführen. https://github.com/distant-viewing/dvt
- **YOLO-Modellfamilie**[3]: Eine frei verfügbare und relativ leicht anpassbare Architektur für praxisnahe Computer-Vision-Anwendungen wie autonomes Fahren, Videoüberwachung und industrielle Qualitätskontrolle. Die Modelle können aber auch sehr gut für andere Aufgaben finegetunt werden und wurde bereits in einem ersten Versuch für das Tracing von Schauspielern in einem Taziyeh-Stück verwendet. https://docs.ultralytics.com/de/
- **ResNet-50 (101)**[4]: Ein von Microsoft Research 2015 veröffentlichtes Modell, das frei

2 https://cinemetrics.uchicago.edu/
3 „You only look once“.
4 https://www.ultralytics.com/de/blog/what-is-resnet-50-and-what-is-its-relevance-in-computer-vision

verfügbar ist und ideal für Transfer Learning sowie zur Erstellung von Embeddings für allgemeine Bilder geeignet ist. Die Modelle wurden auf ImageNet-Daten trainiert, die über 14 Millionen manuell mit über 20.000 Kategorien annotierte Bilder enthalten.

- **FFmpeg**[5]: FFmpeg ist ein freies Kommandozeilenwerkzeug zur Verarbeitung von Audio- und Videodateien, das nahezu alle gängigen Formate und Codecs unterstützt. Es ermöglicht unter anderem das Konvertieren zwischen Videoformaten, das Zuschneiden und Zusammenfügen von Clips, das Extrahieren von Audiostreams sowie das Anpassen von Auflösung, Framerate und Bitrate. In unserer Pipeline spielt FFmpeg eine zentrale Rolle in der Datenvorbereitung: Aus Filmaufnahmen lassen sich damit einzelne Frames als Bilddateien extrahieren. Darüber hinaus kann FFmpeg Videos vor der Verarbeitung auf eine einheitliche Auflösung skalieren, die Framerate reduzieren um Redundanz vermeiden, oder nur relevante Zeitabschnitte ausschneiden.
- **CLIP (Contrastive Language–Image Pretraining)**: CLIP wurde 2021 von OpenAI veröffentlicht und ist ein Modell darauf trainiert, Bilder und Texte in einem gemeinsamen Einbettungsraum zu verorten – es lernt also, welche Textbeschreibung zu welchem Bild „passt". CLIP kann für Zero-Shot-Verfahren genutzt werden, es muss also kein eigenes gelabeltes Dataset aufgebaut werden, sondern kann direkt natürlichsprachige Beschreibungen als Klassifikationskategorien verwenden. Zum Beispiel könnte in einem Bilddatensatz nach einem Bild, das ein Kreuz zeigt, gesucht werden, und CLIP berechnet basierend auf den generierten Embeddings, welches Bild am besten passt. Ein Vorteil von CLIP ist, genauso wie bei YOLO, das CLIP auch lokal betrieben und fine-getunt werden kann.

Konferenzen und Journals

- IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)
- International Journal of Computer Vision (IJCV)
- IEEE Transactions on Image Processing (TIP)
- Computer Vision and Image Understanding (CVIU)

Literatur

Arnold, Taylor, und Lauren Tilton. 2019. „Distant Viewing: Analyzing Large Visual Corpora". *Digital Scholarship in the Humanities* 34 (Supplement_1): i3–16. https://doi.org/10.1093/llc/fqz013.

Arnold, Taylor, und Lauren Tilton. 2023. *Distant Viewing: Computational Exploration of Digital Images*. The MIT Press. https://doi.org/10.7551/mitpress/14046.001.0001.

Arnold, Taylor, Lauren Tilton, und Annie Berke. 2019. „Visual Style in Two Network Era Sitcoms". *Journal of Cultural Analytics* 4 (2): 1194. https://doi.org/10.22148/16.043.

Ayyadevara, V. Kishore. 2024. *Modern Computer Vision with PyTorch: A Practical Roadmap*

5 https://www.ffmpeg.org/

from Deep Learning Fundamentals to Advanced Applications and Generative AI. 1. Aufl. Mit Yeshwanth Reddy. Packt Publishing Limited.

Bamman, David, Rachael Samberg, Richard Jean So, und Naitian Zhou. 2024. „Measuring Diversity in Hollywood through the Large-Scale Computational Analysis of Film“. *Proceedings of the National Academy of Sciences* 121 (46): e2409770121. https://doi.org/10.1073/pnas.2409770121.

Burges, Joel, Nora Dimmock, und Joshua Romphf. 2016. „Collective Reading: Shot Analysis and Data Visualization in the Digital Humanities“. *Cinema Journal Teaching Dossier* 3 (3). https://teachingmedia.org/collective-reading-shot-analysis-and-data-visualization-in-the-digital-humanities/.

Burghardt, Manuel, Michael Kao, und Christian Wolff. 2016. „Beyond Shot Lengths – Using Language Data and Color Information as Additional Parameters for Quantitative Movie Analysis“. *Digital Humanities 2016: Conference Abstracts* (Kraków), 753–55. https://www.researchgate.net/publication/305277308_Beyond_Shot_Lengths_-_Using_Language_Data_and_Color_Information_as_Additional_Parameters_for_Quantitative_Movie_Analysis.

Butler, Jeremy. 2014. „Statistical Analysis of Television Style: What Can Numbers Tell Us about TV Editing?“ *Cinema Journal* 54 (1): 25–44. https://doi.org/10.1353/cj.2014.0066.

Chen, Daoling, und Pengpeng Cheng. 2025. „Digital Extraction and Segmentation of Intangible Cultural Heritage Paper-Cut Patterns“. *Digital Scholarship in the Humanities*, Dezember 16, fqaf006. https://doi.org/10.1093/llc/fqaf006.

Cutting, James E., Kaitlin L. Brunick, Jordan E. DeLong, Catalina Iricinschi, und Ayse Candan. 2011. „Quicker, Faster, Darker: Changes in Hollywood Film over 75 Years“. *I-Perception* 2 (6): 569–76. https://doi.org/10.1068/i0441aap.

Dancygier, Barbara. 2016. „Multimodality and theatre: Material objects, bodies and language“. In *Theatre, performance, and cognition: Languages, bodies and ecologies*, herausgegeben von Rhonda Blair und Amy Cook. Bloomsbury Methuen.

Ferguson, Kevin L. 2016. „Digital Surrealism: Visualizing Walt Disney Animation Studios“. *Digital Humanities Quarterly* 11 (1). https://doi.org/10.63744/xtpq8gu3cfct.

Fischer, Norbert, Dominik Kimmel, und Frank Puppe. 2025. „Semiautomatische Erschließung von Fotografien auf beschrifteten Bildkarten im Archiv. Dokumentenerkennung mit Deep Learning sowie Large-Language-Modellen“. *Zeitschrift für digitale Geisteswissenschaften* 10. https://doi.org/10.17175/2025_009.

Manovitch, Lev. 2020. *Cultural Analytics*. The MIT Press.

Salt, Barry. 1974. „Statistical Style Analysis of Motion Pictures“. *Film Quarterly* 28 (1): 13–22.

https://doi.org/10.2307/1211438.

Schmidt, Thomas, Alina El-Keilany, Johannes Eger, und Sarah Kurek. 2021. *Exploring Computer Vision for Film Analysis: A Case Study for Five Canonical Movies*. https://doi.org/10.5283/EPUB.50867.

Smits, Thomas, und Melvin Wevers. 2023. „A Multimodal Turn in Digital Humanities. Using Contrastive Machine Learning Models to Explore, Enrich, and Analyze Digital Visual Historical Collections“. *Digital Scholarship in the Humanities* 38 (3): 1267–80. https://doi.org/10.1093/llc/fqad008.

Somandepalli, Krishna, Tanaya Guha, Victor R. Martinez, Naveen Kumar, Hartwig Adam, und Shrikanth Narayanan. 2021. „Computational Media Intelligence: Human-Centered Machine Analysis of Media“. *Proceedings of the IEEE* 109 (5): 891–910. https://doi.org/10.1109/JPROC.2020.3047978.

Soriano-Gonzalez, Laura, und Jose Belda-Medina. 2025. „Exploring Image–Text Combinations in Visual Humour through Large Language Models (LLMs)“. *Digital Scholarship in the Humanities* 40 (1): 280–94. https://doi.org/10.1093/llc/fqae068.

Underwood, Ted. 2017. „A Genealogy of Distant Reading“. *Digital Humanities Quarterly* 11 (2). https://doi.org/10.63744/ktwb2atwsbep.