KUNST
CHRONIK

Digital Humanities

Digitale Modalitäten – drei Phasen computationaler Bildbeschreibung

Julian Stalter, M.A.
Doktorand am Institut für Kunstgeschichte und
Wissenschaftlicher Mitarbeiter im Forschungsprojekt
„Reflexionsbasierte künstliche Intelligenz in der
Kunstgeschichte – erklärbare hybride Modelle für die
Bildersuche und -analyse“
Ludwig-Maximilians-Universität München
julian.stalter@kunstgeschichte.uni-muenchen.de

https://doi.org/10.11588/kc.2024.7.105331

77. Jahrgang | Heft 7 | Juli 2024

Digitale Modalitäten – drei Phasen computationaler Bildbeschreibung

Julian Stalter

Ein Arbeitsplatz, aufgenommen aus der Vogelperspektive – eine Hand schiebt von unten einen gelben Post-It-Block ins Bild des Videos. Aus dem Off ist eine menschliche Stimme zu hören, die enthusiastisch fragt: „There we go – tell me what you see," woraufhin eine roboterartige Stimme beginnt, das Setup zu beschreiben. Als die Hand anfängt, Linien auf den Block zu zeichnen, die schließlich in einer abstrakten Form münden, kommentiert die maschinelle Stimme: „The contour lines are smooth and flowing, with no sharp angles or jagged edges." Nun fügt die Hand der Form Füßchen und einen Schnabel hinzu, was mit „It looks like a bird to me" analysiert wird. Es werden wellenförmige Kurven in den Hintergrund gezeichnet, woraufhin die Stimme schlussfolgert: „The bird is swimming in the water […] It is a duck." In dem Video wird die Interaktion von Googles Gemini AI mit Bildern sowie die visuelle Wahrnehmung von Formen und Objekten inszeniert. Der Vorgang demonstriert, wie für Gemini aus der formalen Beschreibung von Linien durch das Hinzufügen von bestimmten Merkmalen ein visuelles Konzept entsteht – die Füße und der Schnabel machen aus den runden Linien einen Vogel. Schließlich wird ein weiteres Bildelement hinzugefügt, das in Relation zu der ersten Form steht – Wellen lassen eine Ente im Wasser schwimmen.

Auch wenn sich dieses Video nachträglich als Zusammenschnitt entpuppt, entsprechen die darin gezeigten visuellen Analysen durch ein multimodales Modell doch dem „state of the art" der computationalen Bilderkennung. Multimodalität wird hier so definiert, dass Datentypen wie Bild, Text und Video gleichzeitig verarbeitet werden können, dass das künstliche neuronale Netz also beispielsweise ein Bild mit Sprache beschreibt. Auch andere Modelle wie BLIP-2 oder GPT-4 erkennen nicht nur bei Zeichnungen, sondern auch, wenn ihnen ein Kunstwerk vorgelegt wird, dessen Komposition, Aufbau und Inhalt und liefern Deskriptionen, die ikonographische und ikonologische Aspekte mit einbeziehen. Es sind Beschreibungen, die nicht als tiefgreifende kunsthistorische Erkenntnis verstanden werden sollten, aber doch als algorithmische Verarbeitung eines visuellen „Inputs", an dessen Ende die Produktion eines Textes steht. „Erst in der Sprache gewinnt der Wissenschaftler eine Instanz, die zur Kontrolle und Kritik seiner Einsichten geeignet ist", sagte Gottfried Boehm über die erkenntniskritische Rolle der Bildbeschreibung (Boehm 1995, 24). Die Sprache, die einer Bildbeschreibung zugrunde liegt, ist kein leicht operationalisierbarer Prozess, sondern abhängig vom Kontext und der Interpretation des Betrachters. Und so mag auch das Ergebnis dieser Modelle vielleicht trivial erscheinen, es eröffnet aber doch ein Feld für die kritische Auseinandersetzung mit den zugrundeliegenden Prämissen dieses „Outputs".

Der Kunsthistoriker und der Computer

Etwa 60 Jahre bevor das erwähnte Video entstand, beschrieb der Kunsthistoriker Jules Prown in seinem Aufsatz „The Art Historian and the Computer" verschiedene Methoden und Fähigkeiten von Forscher:innen im Vergleich zu denen von Computern. Ihm zufolge widmen sich Kunsthistoriker:innen den qualitativen Differenzierungen, die nur mittels eines geschulten Verstandes und eines sensiblen Auges möglich seien. Der Computer operiere hinge-

gen mit quantitativen Berechnungen und einer beeindruckenden Geschwindigkeit, die jedoch nach Prown durch ihre monotone und unreflektierte Effizienz jedem anständigen Kunsthistoriker einen Schauer über den Rücken jage (Prown 1966). Dies verstand Prown aber nicht als Ausschlusskriterium, sondern er nutzte den Computer ausgiebig für die eigene Forschung.

Seit Prowns Zeiten, als erste Rasterungsalgorithmen genutzt wurden, hat sich die digitale Bildanalyse weiterentwickelt. Fortschritte im Bereich der Computertechnik und neue Zielsetzungen haben zu einer Vielzahl von Ansätzen in der digitalen Bildanalyse geführt, bei denen jeweils spezifische Methoden zum Einsatz kommen. Im Folgenden werden drei Phasen der computergestützten Bildbeschreibung vorgestellt. Dabei wird erörtert, wie sie unter den jeweiligen technischen und methodischen Gegebenheiten definiert werden, welche Paradigmen sich daraus ergeben und welche Kritikpunkte auftreten. Denn die von Prown behauptete Trennung zwischen qualitativen Differenzierungen und quantitativen Berechnungen hat über die Jahrzehnte nicht Bestand gehabt. Vielmehr haben sich Verbindungen entwickelt, welche die vermeintlich strikt getrennten Charakteristika ineinander übergehen und in beide Richtungen wirken lassen. Abschließend gilt es zu erörtern, welche Auswirkungen die vorgestellten Methoden der Bildbeschreibung auf die Prozesse kunsthistorischer Forschung haben und welche neuen Herausforderungen sowie mögliche Forschungsgebiete sich daraus ergeben.

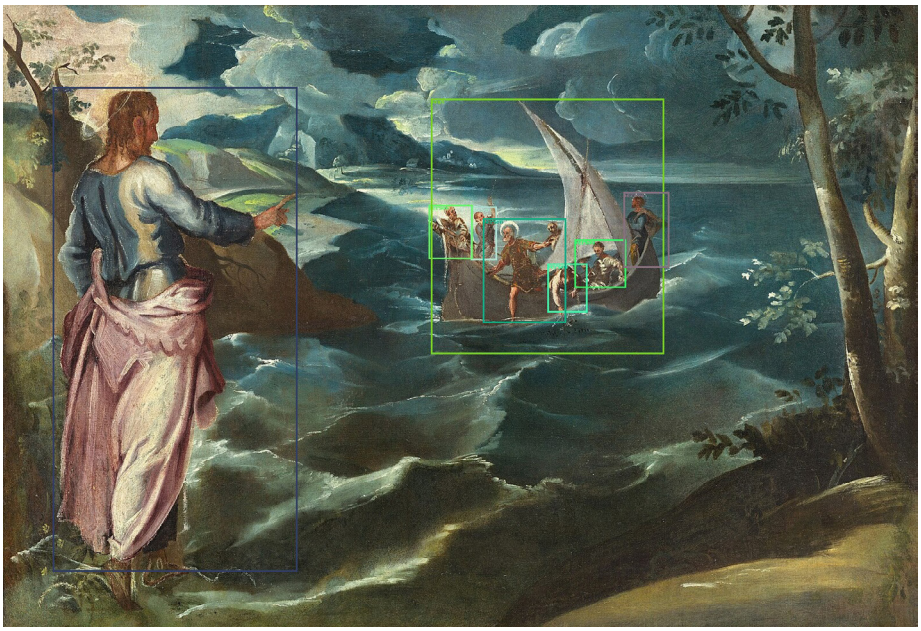
Elektronische Kunstbotanik? Digitale Dokumentationssysteme

Bevor größere kunsthistorische Bildbestände systematisch digitalisiert wurden, lag der Fokus auf textbasierten Informationssystemen, die eine schnelle Abfrage von Daten und Schlagworten ermöglichten. Erstmals 1983 auf dem Kunsthistorikertag in Wien vorgestellt, wurden die neuen digitalen Systeme ambitioniert präsentiert, aber auch skeptisch aufgenommen. Lutz Heusinger, damals Leiter von Foto Marburg und Initiator des Dokumentationssystems MIDAS, diskutierte die Entwicklungen und veröffentlichte kurz darauf acht Thesen zu Kunstgeschichte und elektronischer Datenverarbeitung in den *kritischen berichten* (Heusinger 1983, 67). Darin stellt er die Beschreibung von Kunstwerken mittels eines durch digitale Datenbanken angereicherten Forschungsprozesses vor, der es ermöglichen soll, kunsthistorische Sachverhalte elektronisch zu ordnen und zu verarbeiten. So sollen Daten zu Kunstwerken, Künstler:innen, Ereignissen, Textquellen, aber auch Bildinhalten erfasst werden. Letzteres mit Hilfe von Iconclass, einem Klassifikationssystem für deren Erschließung (vgl. hierzu den Beitrag von Berthold Kreß in: *Kunstchronik* 77/2, 2024, 128–140, DOI↗).

Dieses Dokumentationssystem wirft jedoch, wie Heusinger selbst bemerkt, spezifische Probleme auf. So würden wir am Ende dieser Entwicklung zwar jedes Werk in Sekundenschnelle finden, aber die Objekte werden uns aus ihrem Überlieferungskontext gerissen gegenüberstehen. Damit formuliert Heusinger eine Kritik, mit der sich die digitale Kunstgeschichte auch in der Folge häufig konfrontiert sah und sieht: Der Entstehungskontext der Forschungsliteratur mit ihren gesellschaftlichen Prämissen werde zu wenig berücksichtigt – eine Kritik, die auch in Reaktion auf Heusingers Vortrag geäußert wurde. Karl Clausberg veröffentlichte eine Glosse mit dem Titel „1984 wieder hinter Schloss(er) und Riegl? Von der Wiener Formengeschichte zur elektronischen Kunstbotanik“. Er verstand die formalisierte Verarbeitung als „erkennungsdienstliche Behandlung“ der Kunstgeschichte und sah in der digitalen Klassifikation einen disziplingefährdenden Katarakt (Clausberg 1983, 71). Die Idee der Beschreibung des Werkes in Form der formalen Verschlagwortung, obgleich in Bibliothekskatalogen längst üblich, führte bei der digitalen Bildanalyse zu Alarmismus, zumal in einem Umfeld, das damals Kunstwerke sozialgeschichtlich kontextualisierte. So prophezeite Clausberg, wie am Titel seiner Glosse ablesbar, einen Rückfall hinter die Protagonisten der Wiener Schule und soziologisch-kulturhistorische Denkmodelle in eine morphologische Beschreibung und Kategorisierung (Pratschke 2016, 60).

Arrays auslesen: Computergestützte Bildanalyse

Die extrinsische Beschreibung eines Werkes mit Hilfe von Datierung, Schriftquellen oder ikonographischer Codierung wurde in den 1990er Jahren durch intrinsische Beschreibungen nach formalen Kriterien ergänzt. Da Bilder in digitaler Form in erster Linie Aneinanderreihungen von Zahlenfeldern sind, die vom Computer gelesen und in Pixel übersetzt werden, ermöglichen neue Methoden das Auslesen dieser Arrays. Dieser Prozess bietet für die Kunstgeschichte nun die Möglichkeit, mit den Inhalten, wenn auch auf abstrakter Ebene, zu interagieren. Durch computergestützte Bildanalyse kann so auf formale Eigenheiten von Kunstwerken zugegriffen werden. In der Linguistik hatten sich ähnliche Methoden bereits früher etabliert, die Kunstgeschichte sah sich jedoch vor größere Herausforderungen gestellt. Wie William Vaughan beschreibt, bestehen Bilder eben nicht aus diskreten und leicht identifizierbaren Elementen wie etwa Sätzen (Vaughan 1997, 97). Durch technische Fortschritte bot der computergestützte formalanalytische Ansatz nun jedoch die Möglichkeit, visuelle Eigenschaften wie die Länge und Richtung einer Linie, die Dichte eines Farbtons und dessen Deckung sowie die Farbe als solche zu bestimmen. In ersten Anwendungen sind Programme wie Morelli in der Lage, verschiedene Reproduktionen eines identischen Kunstwerkes zu erkennen. Der Name des Programms ist eine Hommage an Giovanni Morelli, der im 19. Jahrhundert anhand nebensächlicher Details wie Händen und Ohren kennerschaftliche Zuschreibungen vornahm. Vaughan, der der klassischen Formanalyse aus ihrer Abseitsstellung heraushelfen wollte, stellte mit dem Programm zumindest theoretisch eine Möglichkeit zur Verfügung, Kopien, Fälschungen oder Abhängigkeitsverhältnisse zwischen Bildern zu rekonstruieren (Kohle 2013, 49). Das geschulte Auge der Kunsthistoriker:innen als genuin menschliche Kompetenz, von dem noch Prown ausging, wird nun vor allem unter dem Aspekt der Kennerschaft, die sich auf den Computer bezieht, herausgefordert (vgl. den Beitrag von Christine Tauber in Teil II dies *Special Issue*). Seit 2015 verbessern künstliche neuronale Netze die Bildanalyse am Computer durch Techniken wie Objekterkennung und Bildsegmentierung. Häufig kommen Deep Convolutional Neural Networks (CNNs) zum Einsatz und „lernen", Objekte und Personen in Bildern zu klassifizieren. | Abb. 1 | Bei der „Feature Extraction" werden visuelle Merkmale extrahiert und



| Abb. 1 | Personen- und Objekterkennung in Jacopo Tintorettos „Christus am See Tiberias" mit DEtectionTRansformer. Graphik: Julian Stalter

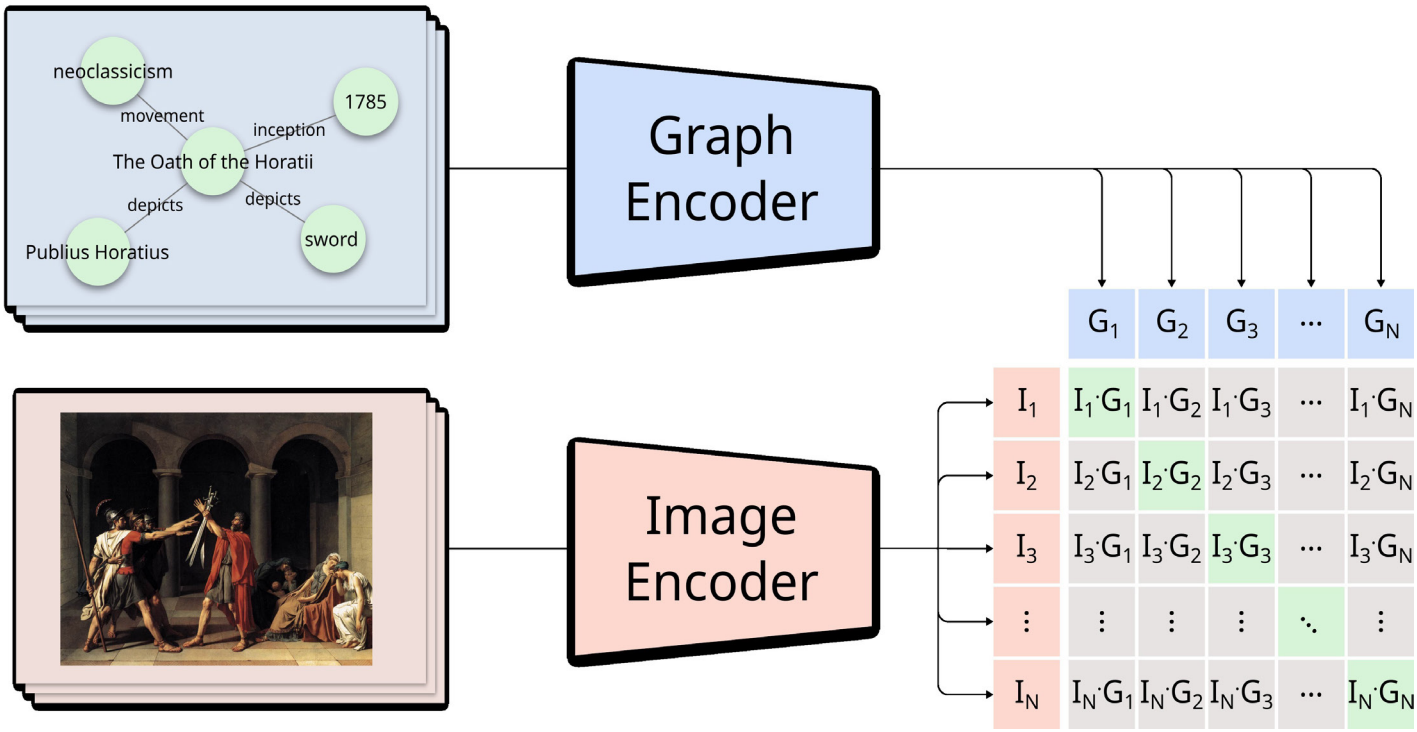
anschließend das Modell mithilfe von Menschen annotierter Bilder trainiert. So können Objekte und Figuren klassifiziert werden. Dies ermöglicht das Durchsuchen und Ordnen großer kunsthistorischer Datensätze, beispielsweise nach dargestellten Posen (Schneider 2024).

Amanda Wasielewski, die diese Entwicklungen unter dem Stichwort des „computational formalism" behandelt, konstatiert hierbei eine Rückkehr formalistischer Methoden. Die Analyse der Merkmale, die nunmehr verschärfte Ausklammerung des Kontextes, der Vergleich mit anderen Werken und die Klassifizierung von Kunstwerken seien einer rein formalen Herangehensweise sehr ähnlich. Dennoch unterscheidet sie zwei neue Ausprägungen des „computational formalism" von der traditionellen Methodik. Zum einen werde der Computer nun zur Analyse großer Datenmengen eingesetzt, zum anderen nehme ein neuronales Netz Merkmale anders wahr als ein Kunsthistoriker (Wasielewski 2023, 32ff.). Die Wahrnehmung durch ein Modell stellt kein ‚Sehen' im menschlichen Sinne von Objekten und Relationen im Bild dar, sondern ist innerhalb der Black-Box des neuronalen Netzes schwer nachvollziehbar und muss kritisch reflektiert werden. Diesen *Perceptual Bias* gilt es bei der Anwendung dieser Methoden zu berücksichtigen und auch die Visualisierung von extrahierten Merkmalen, um beispielsweise Modellklassifikationen nachzuvollziehen, sollte als *technical metapicture* verstanden werden (Bell/Offert 2021, 1133).

Text-Bild-Paare: Multimodale Modelle

Das Beispiel von Googles Gemini AI, bei dem eine Stimme visuellen Input beschreibt, verdeutlicht die Fähigkeiten neuer multimodaler Modelle zur gemeinsamen Verarbeitung von Text, Video und Bild. Während CNNs mit Bildern und annotierten Kategorien trainiert wurden, werden multimodale Modelle mit Bild-Text-Paaren gespeist. | Abb. 2 | Dies sind zumeist Bilder und Bildbeschreibungen aus dem Internet. Das bekannteste Modell, CLIP, erschien 2021 und wurde mit 400 Millionen Bild-Text-Paaren trainiert. In dem multimodalen Modell werden diese Paare zunächst codiert und anschließend als Vektoren in einem gemeinsamen Merkmalsraum abgebildet („embedded"). Dieser kann mehrere hundert Dimensionen umfassen – wobei dort gemeinsame Merkmale beider Modalitäten räumlich repräsentiert werden – und ermöglicht es, komplexe Zusammenhänge und Ähnlichkeiten zwischen den Paaren zu erkennen. Während des Trainingsprozesses werden die Merkmale so angelegt, dass sich ähnliche Bild-Text-Elemente in räumlicher Nähe zueinander befinden. Dies ermöglicht es dem Modell, relevante Verknüpfungen zwischen Bildern und Texten effizient zu erkennen und abzufragen.

Für die Kunstgeschichte stellt diese Verschränkung von Bild und Text eine interessante Entwicklung dar. Die Objekte der Kunstgeschichte sind durchgängig von Texten geprägt: Die Bilder werden von der Verfasstheit unserer Texte konzeptualisiert, und die Texte der Kunstgeschichte sind entlang von Bildern konzipiertes Wissen (vgl. Pias 2003, 23). Durch die multimodale Art des Trainings tritt in den genannten Modellen der (Kon-)Text stärker hervor. In eine Suchmaschine implementiert, kann CLIP nicht nur Objekte erkennen, sondern findet beispielsweise auch abstrakte visuelle Konzepte zu komplexen Anfragen wie „Sommer" oder „Rhythmus" (Impett/Offert 2023, 7). Grundlegend für die Qualität dieser Modelle sind die Trainingsdaten, und hier bietet sich für die Kunstgeschichte die Chance, die Entwicklung aktiv zu verbessern und mitzugestalten. Ähnlichkeitssuchen, die Suche nach Bildern mit natürlichsprachigem Text und das Generieren von Bildbeschreibungen werden durch diese Modelle ermöglicht – und optimiert, wenn entsprechend hochwertige Datensätze aus der Kunstgeschichte als Texte oder Wissensgraphen bereitgestellt werden (Stalter/Springstein 2024). Durch die reflexive Betrachtung der Auswirkungen des Trainings und des Einsatzes von multimodalen Modellen könnten interessante Forschungsfragen entstehen, die nun nicht mehr nur auf Materialselektion zielen, sondern ihrerseits Methoden und Quellen reflektieren: Wie unterscheiden sich Bildbeschreibungen



| Abb. 2 | Training eines multimodalen Modells mit Informationen aus einem kunsthistorischen Wissensgraphen. Graphik: Matthias Springstein

eines Modells, das mit Texten der Kunstkritik aus dem 19. Jahrhundert trainiert wurde, von einem Modell, das Aufsätze des 21. Jahrhunderts verarbeitet hat? Dies ermöglicht nicht nur die Analyse der epistemologischen Prämissen und Charakteristika der Forschungsdaten, sondern erlaubt auch eine methodische Reflexion darüber, wie die Modelle diese Informationen verarbeiten. Die Kunsthistoriker:innen operieren hierbei in einer kuratorischen Position: Sie stellen die Trainingsdaten zusammen und analysieren kritisch die Ergebnisse. Sie sollten nicht, wie Prown es dem Computer vorwirft, in gedankenloser Effizienz möglichst viel Material sammeln, sondern je nach Forschungsfrage und Methodik anhand ihrer Expertise auswählen.

Perspektiven

Eingangs wurde Gottfried Boehm zitiert, der die Bildbeschreibung als Instanz zur Kontrolle und Kritik der wissenschaftlichen Einsichten betrachtet. Wie anhand der drei Entwicklungsphasen gezeigt, bringen die verschiedenen Methoden je eigene Modi hervor, die mit Limitationen, aber auch neuen Möglichkeiten der Analyse des Kunstwerks und von dessen Kontext verbunden sind. Gerade wegen dieser Fortschritte bleiben die Kunsthistoriker:innen unverzichtbar: Sie müssen mit ihrer Expertise diese Prozesse begleiten und formen. Das Ordnen und Maschinenlesbarmachen von Trainingsdaten mag auf den ersten Blick mühsam und wenig anspruchsvoll erscheinen, stellt jedoch sowohl technisch als auch intellektuell eine erhebliche Herausforderung dar. Es gilt, eine gewisse Komplexität zu bewahren, während gleichzeitig eine notwendige Formalisierung erfolgt. Auch die Ergebnisse dieser Prozesse müssen kritisch analysiert und evaluiert werden. Es ist wichtig, über alternative Formen der qualitativen Bewertung und ihrer Kriterien nachzudenken, die sich von den häufig quantitativ ausgerichteten Methoden in der Informatik abheben. Heat Maps ausschlaggebender Bildbereiche und die Berechnung von Bild-Text-Ähnlichkeiten bieten Anhaltspunkte für Klassifikationen und Retrieval-Ergebnisse der Modelle und können von Kunsthistoriker:innen fallspezifisch analysiert werden. Auch eine qualitative Analyse der extrahierten Trainingsdaten, beispielsweise in einem Wissensgraph, bietet mehr Transparenz bei der Anwendung neuronaler Netze. Diese Überlegungen böten auch die Möglichkeit, Skeptiker:innen stärker einzubinden und digitale Methoden als selbstverständlichen Bestandteil in die kunsthistorische Disziplin zu integrieren.

Literatur

Bell/Offert 2021: Peter Bell und Fabian Offert, Perceptual bias and technical metapictures: critical machine vision as a humanities challenge, in: AI & Society 36, 2021, 1133–1144.

Bell/Ommer 2016: Peter Bell und Björn Ommer, Digital connoisseur? How computer vision supports art history, in: Stefan Albl und Alina Aggujaro (Hg.), Il metodo del conoscitore. Approcci, limiti, prospettive. Connoisseurship nel XXI secolo, Rom 2016, 187–200.

Boehm 1995: Gottfried Boehm, Bildbeschreibung. Über die Grenzen von Bild und Sprache, in: Gottfried Boehm und Helmut Pfotenhauer (Hg.), Beschreibungskunst – Kunstbeschreibung, München 1995, 23–40.

Clausberg 1983: Karl Clausberg, 1984 wieder hinter Schloss(er) und Riegl? Von der Wiener Formengeschichte zur elektronischen Kunstbotanik, in: kritische Berichte 11/3, 1983, 71–74.

Heusinger 1983: Lutz Heusinger, Kunstgeschichte und EDV: 8 Thesen, in: kritische berichte 11/4, 1983, 67–70.↗

Impett/Offert 2023: Leonardo Impett und Fabian Offert, There is a Digital Art History, Arxiv 2023.↗

Kohle 2013: Hubertus Kohle, Digitale Bildwissenschaft, Glückstadt 2013.

Lang/Ommer 2021: Sabine Lang und Björn Ommer, Transforming Information Into Knowledge: How Computational Methods Reshape Art History, in: Digital Humanities Quaterly (DHQ) 15/3, 2021.↗

Pias 2003: Claus Pias, Das digitale Bild gibt es nicht. Über das (Nicht-)Wissen der Bilder und die informatische Illusion, in: zeitenblicke 2/1, 2003.

Pratschke 2016: Margarete Pratschke, Wie Erwin Panofsky die Digital Humanities erfand. Für eine Geschichte und Kritik digitaler Kunst- und Bildgeschichte, in: kritische berichte 44/3, 2016, 56–66.

Prown 1966: Jules Prown, The Art Historian and the Computer. An Analysis of Copley's Patronage, 1753–1774, Yale Archive 1966.

Raspe/Schelbert 2024: Martin Raspe und Georg Schelbert, Bilder ohne Worte? Kunstgeschichte auf dem Weg in die praktische Digitalität, in: Lisa Dieckmann u. a. (Hg.), 4D: Dimensionen | Disziplinen | Digitalität | Daten (Computing in Art and Architecture, 6), Heidelberg: arthistoricum.net-ART-Books, 2022 (2024).↗

Schneider 2024: Stefanie Schneider, My Body is a Cage. Human Pose Estimation und Retrieval in kunsthistorischen Inventaren, DHd 2024, Quo Vadis DH?, 231–236.↗

Stalter/Springstein 2024: Julian Stalter, Matthias Springstein u. a., ReflectAI: Reflexionsbasierte künstliche Intelligenz in der Kunstgeschichte, DHd 2024. Quo Vadis DH?, 414–417.↗

Vaughan 1997: William Vaughan, Computergestützte Bildrecherche und Bildanalyse, in: Hubertus Kohle (Hg.), Kunstgeschichte digital, Berlin 1997, 97–105.

Wasielewski 2023: Amanda Wasielewski, Computational Formalism. Art History and Machine Learning, Cambridge/London 2023.