

ReflectAI: Reflexionsbasierte künstliche Intelligenz in der Kunstgeschichte

Stalter, Julian

julian.stalter@kunstgeschichte.uni-muenchen.de
Ludwig-Maximilians-Universität München, Deutschland
ORCID: 0000-0003-1149-1688

Springstein, Matthias

Matthias.Springstein@tib.eu
L3S Research Center, Leibniz Universität Hannover,
Deutschland; TIB – Leibniz-Informationszentrum
Technik und Naturwissenschaften, Deutschland
ORCID: 0000-0002-6509-8534

Kristen, Maximilian

max.kristen@campus.lmu.de
Ludwig-Maximilians-Universität München, Deutschland

Schneider, Stefanie

stefanie.schneider@itg.uni-muenchen.de
Ludwig-Maximilians-Universität München, Deutschland
ORCID: 0000-0003-4915-6949

Müller-Budack, Eric

Eric.Mueller@tib.eu
L3S Research Center, Leibniz Universität Hannover,
Deutschland; TIB – Leibniz-Informationszentrum
Technik und Naturwissenschaften, Deutschland
ORCID: 0000-0002-6802-1241

Ewerth, Ralph

Ralph.Ewerth@tib.eu
L3S Research Center, Leibniz Universität Hannover,
Deutschland; TIB – Leibniz-Informationszentrum
Technik und Naturwissenschaften, Deutschland
ORCID: 0000-0003-0918-6297

Kohle, Hubertus

hubertus.kohle@lmu.de
Ludwig-Maximilians-Universität München, Deutschland
ORCID: 0000-0003-3162-1304

In der Kunstgeschichte sind Ähnlichkeitsbewertungen von Bildern von großer Bedeutung: Wölfflin analysierte Kunstwerke nach begrifflichen Gegensatzpaaren mit Doppelprojektionen, Warburg beim *Vergleichenden Sehen*

nach sogenannten *Pathosformeln* (Wölfflin, 1915; Hensel, 2010). Mit Verfahren aus dem Bereich des maschinellen Sehens (*Computer Vision*) lassen sich derartige Bewertungen heute automatisieren. *State-of-the-Art*-Modelle wie *GPT* (*Generative Pre-trained Transformer*; OpenAI, 2023) oder *CLIP* (*Contrastive Language-image Pre-training*; Radford et al., 2021) können zudem aufgrund erweiterter Rechenkapazitäten effizienter generalisieren und damit auf nicht realweltliche Bilddaten angewandt werden. Eingesetzt wurden solche Ansätze bereits in kunsthistorischen Projekten, insbesondere in der Ähnlichkeits-basierenden Bildsuche und -clustering (Schneider et al., 2022; Offert und Bell, 2023).

Der Einsatz von *Künstlicher Intelligenz* (KI) in der kunsthistorischen Forschung eröffnet unbestreitbar neue explorative Potenziale für die Analyse von Bildähnlichkeiten. Aus methodenkritischer Perspektive ist er jedoch zu hinterfragen; insbesondere der „Black Box“-Charakter künstlicher neuronaler Netze wird diskutiert (Crawford und Paglen, 2021). Diesen Problemen soll sich das hier vorgestellte Projekt *ReflectAI* speziell für den Bereich der Kunstgeschichte annehmen. An dem Vorhaben, das im Rahmen des DFG-Schwerpunktprogramms „Das digitale Bild“ von 2022 bis 2025 gefördert wird, sind Forschende aus der Kunstgeschichte und Informatik der Universitäten München und Hannover beteiligt. Eine schematische Darstellung aller Teilmodule innerhalb des Projekts ist in Abb. 1 dargestellt.

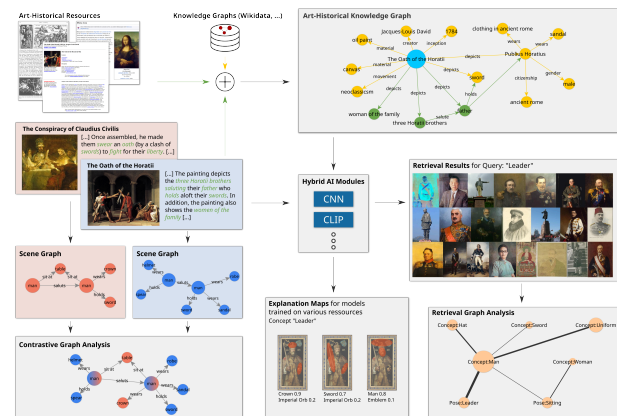


Abb. 1: Schematische Darstellung der Erstellung kunsthistorischer Wissensgraphen auf der Basis domänenspezifischer Textkorpora, des Trainings multimodaler Modelle sowie der Visualisierungstypen, die im Rahmen von *ReflectAI* entwickelt werden.

Herausforderung „Black Box“

Obwohl die Ein- und Ausgabedaten neuronaler Netze bekannt sind, bleiben die internen Verarbeitungsprozesse weitgehend undurchsichtig. Ziel unseres Projekts ist es, ein tieferes Verständnis der Prozesse zu erlangen, die zu den Suchergebnissen führen. Dabei konzentrieren wir uns auf Prozesse, die zur Identifikation von Bildähnlichkeiten, zum

Clustering von Bildern und zur Klassifikation anhand spezifischer Bilddetails führen. Im Sinne einer „Explainable AI“ werden Methoden untersucht, die einen Blick in diese „Black Box“ erlauben (Guidotti et al. 2019; Offert und Bell, 2021). Unsere Absicht ist, Verfahren der automatisierten Bildanalyse sowohl aus bildwissenschaftlicher als auch aus wissenschaftshistorischer Perspektive zu analysieren und in eine für kunsthistorische Forschungsprojekte optimierte Anwendung zu überführen. Mit Motivation entwickeln wir im Projekt reflexive Werkzeuge und stellen sie bereit.

Reflexive Komponenten

Expertenwissen

Die Zusammensetzung der Trainingsdaten, auf deren Basis neuronale Netze und Modelle trainiert werden, ist den Anwendern oft nicht bekannt und im Sinne einer Methodenkritik schwer zugänglich. Unser Projekt beinhaltet die Optimierung dieser Modelle mit kunsthistorischen Texten und multimodalen Informationen (wie Text-Bild-Paaren), um domänenspezifisches Wissen zu integrieren. Dabei wollen wir z.B. auf Visual Language Models wie BLIP-2 (Li, 2023) oder LLaVA (Liu, 2023) zurückgreifen, wobei das Expertenwissen als Trainingsmaterial für Modelle dienen kann oder auch als zusätzlicher Input des Language Models genutzt werden kann. Außerdem trainieren wir Modelle mit Textkorpora: So können beispielsweise die Ergebnisse von Modellen, in die Kunstkritiken des 19. Jahrhunderts eingespeist wurden, mit denen verglichen werden, die auf Texten des 21. Jahrhunderts basieren. Im Rahmen des Projekts wird dazu ein multimodales Datenkorpus aus Bild- und Textquellen aufgebaut. Es umfasst derzeit 60.000 Objektbeschreibungen aus Museen, Sammlungen und Auktionshäusern, 10.000 Lexikonartikel und 50.000 wissenschaftliche Publikationen.

Wissensgraphen

Das so gesammelte Expertenwissen soll nicht nur zum Training der Modelle, sondern auch zur Erstellung von *Wissensgraphen* (*Knowledge Graphs*) verwendet werden. Obwohl es Versuche gibt, kunsthistorisches Wissen aus Wissensgraphen für das Training neuronaler Netze zu nutzen, beschränken sich diese häufig auf einzelne kunsthistorische Attribute (Garcia und Vogiatzis, 2018; Schrade und Söhn, 2022). Neben den Wissensgraphen werden Szenegraphen generiert, die mit zusätzlichen Visualisierungen ein besseres Verständnis der Bildauswahl anhand der in den Suchergebnissen dargestellten Objekte und Beziehungen ermöglichen und auf Verzerrungen („Biases“) oder Ähnlichkeiten hinweisen können (Suhail, 2021). Diese Szenegraphen können Beziehungen zwischen Objekten innerhalb eines Kunstwerks oder im Kontext darstellen. In Kombination mit Wissensgraphen ist eine kontrastive Analyse möglich.

Visualisierung von Bias

Unter Bias versteht man die Verzerrung der Ergebnisse aufgrund falscher Annahmen über die Trainingsdaten oder die Modelle. Dieser Bias kann historisch bedingt sein oder durch fehlerhafte oder diskriminierende Annotationen in den Trainingsdaten entstehen (Pasquinelli und Joler, 2021). Um solche Verzerrungen sichtbar zu machen, werden den Nutzenden im Rahmen des Projekts verschiedene Werkzeuge zur Verfügung gestellt, mit denen Biases visualisiert werden können; ein Beispiel ist in Abb. 2 dargestellt.

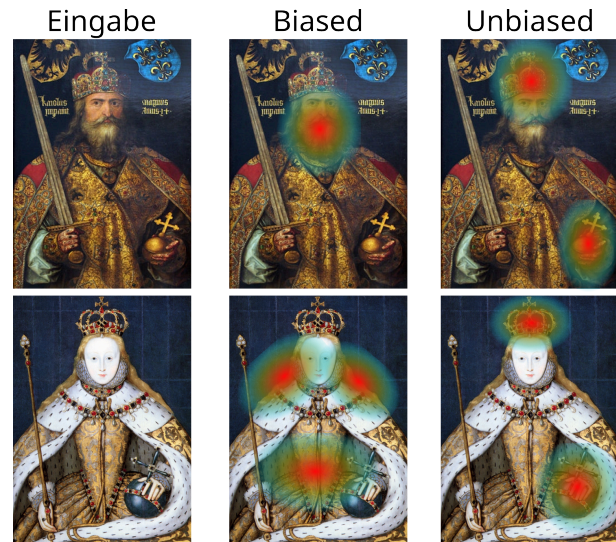


Abb. 2: Illustration der Vorhersage von Modellen zur Erkennung von herrschenden Personen. In verzerrten Modellen werden geschlechtsspezifische Merkmale wie der Bart oder das lange Haar zur Vorhersage verwendet, in unverzerrten Modellen eher Krone und Reichsapfel.

Danksagung

Das Projekt wird von der Deutschen Forschungsgemeinschaft (DFG) unter der Projektnummer 510048106 gefördert.

Bibliographie

- Crawford, Kate und Trevor Paglen.** 2021. „Correction to: Excavating AI. The Politics of Images in Machine Learning Training Sets.“ *AI & Society* 36.4: 1399 10.1007/s00146-021-01301-1.
- Garcia, Noa und George Vogiatzis.** 2018. „How to Read Paintings. Semantic Art Understanding with Multimodal Retrieval.“ In *Computer Vision – ECCV 2018 Workshops. Lecture Notes in Computer Science* 11130: 676–691 10.1007/978-3-030-11012-3_52.
- Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti und Dino Pedreschi.** 2019. „A Survey of Methods for Explaining

Black Box Models.“ In *ACM Computing Surveys* 51.5: 93:1–93:42 10.1145/3236009.

Hensel, Thomas. 2010. „Aby Warburg und die ‚Verschmelzende Vergleichsform‘.“ In *Vergleichendes Sehen*, hg. von Lena Bader, Martin Gaier und Falk Wolf, 468–489. München: Fink.

Li, Junnan, Dongxu Li, Silvio Savarese und Steven C. H. Hoi 2023. „BLIP-2: Bootstrapping Language-image Pre-training with Frozen Image Encoders and Large Language Models.“ In *International Conference on Machine Learning. ICML 2023*, 19730–19742. <https://proceedings.mlr.press/v202/li23q.html> (zugegriffen: 3. Dezember 2023).

Liu, Haotian, Chunyuan Li, Qingyang Wu und Yong Jae Lee. 2023. *Visual Instruction Tuning*. *arXiv:2304.08485*.

Offert, Fabian und Peter Bell. 2023. „imgs.ai. A Deep Visual Search Engine for Digital Art History.“ In *DH 2023. Digital Humanities 2023. Conference Abstracts*, 141–142 10.5281/zenodo.7961822.

Offert, Fabian und Peter Bell. „Perceptual Bias and Technical Metapictures. Critical Machine Vision as a Humanities Challenge.“ *AI & Society* 36.4: 1133–1144 10.1007/s00146-020-01058-z.

OpenAI. 2023. *GPT-4 Technical Report*. *arXiv:2303.08774*.

Pasquinelli, Matteo und Vladan Joler. 2021. „The Noosope Manifested. AI as Instrument of Knowledge Extractivism.“ *AI & Society* 36.4: 1263–1280 10.1007/s00146-020-01097-6.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger und Ilya Sutskever. 2021. „Learning Transferable Visual Models From Natural Language Supervision.“ In *Proceedings of the 38th International Conference on Machine Learning. ICML 2021*, hg. von Marina Meila und Tong Zhang, 8748–8763, <http://proceedings.mlr.press/v139/radford21a.html> (zugegriffen: 19. Juli 2023).

Schneider, Stefanie, Matthias Springstein, Javad Rahnama, Hubertus Hohle, Ralph Ewerth und Eyke Hüllermeier. 2022. „iART. Eine Suchmaschine zur Unterstützung von bildorientierten Forschungsprozessen.“ In *8. Tagung des Verbands Digital Humanities im deutschsprachigen Raum. DHd 2022*, hg. von Michaela Geierhos, 142–147 10.5281/zenodo.6304590.

Schrade, Torsten, und Linnaea Söhn. 2022. *Culture Portal 1.3. Knowledge Graph v.1.0 & Repositorien Überblick*. <https://nfdi4culture.de/de/nachrichten/culture-portal-13-knowledge-graph-v10-repository-overview.html> (zugegriffen: 19. Juli 2023).

Suhai, Mohammedl, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gérard G. Medioni und Leonid Sigal. 2021. „Energy-Based Learning for Scene Graph Generation.“ In *IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2021*, 13936–13945 10.1109/CVPR46437.2021.01372.

Wölfflin, Heinrich. 1915. *Kunstgeschichtliche Grundbegriffe. Das Problem der Stilentwicklung in der neueren Kunst*. München: F. Bruckmann.