# Computationale Beschreibung visuellen Materials am Beispiel des Graphic Narrative Corpus

**Laubrock, Jochen**
laubrock@uni-potsdam.de
Universität Potsdam, Deutschland

**Dubray, David**
ddubray@uni-potsdam.de
Universität Potsdam, Deutschland

**Krügel, André**
kruegel@uni-potsdam.de
Universität Potsdam, Deutschland

## Einleitung

Die digitale Revolution in den Geisteswissenschaften hat diesen eine Reihe neuer Methoden eröffnet. Durch die heute verfügbaren großen Datenmengen und intelligenten Algorithmen haben sich zwar einige bisher offengeliebene geisteswissenschaftliche Kernfragen beantworten lassen, jedoch wird mancherorts eine Methodenfokussiertheit und Theoriemangel kritisiert (Gumbrecht, 2014). Ob die Digitalen Geisteswissenschaften zu einer darüber hinausgehenden tiefergreifenden Veränderung der geisteswissenschaftlichen Erkenntnis führen werden, bleibt abzuwarten; derzeit scheint es, als werde das Potenzial der neuen methodischen Zugänge erst noch ausgelotet. In anderen Wissenschaften hat aber die Verfügbarkeit computationaler Modelle und der damit einhergehende Zwang, implizite Annahmen zu explizieren und Theorien formal testbar zu machen, signifikant zur Theoriebildung und -prüfung beigetragen (cf. Myung & Pitt, 2002; Lewandowsky & Farrell, 2011). Deshalb besteht die begründete Hoffnung, dass computationale Modellierung auch die Geisteswissenschaften bereichern wird.

Ein Großteil der Forschung in den digitalen Geisteswissenschaften beschäftigt sich mit Text. Es gibt hier eine fruchtbare interdisziplinäre Zusammenarbeit von Literaturwissenschaften und Computerlinguistik; im Umfeld des "Distant Reading" sind umfangreiche Werkzeuge entstanden, mit denen sich etwa stilometrische Analysen oder Topic Modeling computergestützt vornehmen lassen (Blei, 2012; Juola, 2006). Auch netzwerkanalytische Methoden aus der theoretischen Physik und computationalen Soziologie haben hier interessante neue Perspektiven eröffnet (Schich et al., 2014). Dagegen ist die digitale Analyse visuellen Materials noch relativ wenig entwickelt oder standardisiert, obwohl dieses für Disziplinen wie z.B. Kunstgeschichte oder Archäologie von zentralem Interesse ist. In den letzten Jahren wurden durch Entwicklungen im Bereich der Convolutional Neural Networks (CNN) die Möglichkeiten automatisierter Bildanalyse revolutioniert. Während in klassischen Ansätzen der maschinellen Bildverarbeitung ein hohes Ausmaß an Expertenwissen notwendig war, um Merkmale zu definieren, mit denen sich das Material sinnvoll beschreiben ließ ("engineered features"), lernen CNNs die Merkmale durch Fehlerrückführung (Backpropagation) selbst.

Convolutional Neural Networks sind eine besondere Klasse künstlicher neuronaler Netze, die sich durch eine 2D-Anordnung der Neuronen, innerhalb einer Schicht geteilte Gewichte und lokale Konnektivität auszeichnen. Sie eignen sich insbesondere für die Analyse von Bildmaterial. Die Netze sind typischerweise auf einer großen Anzahl von Fotos in Objektklassifikationsaufgaben trainiert worden, dabei bilden sich auf verschiedenen Ebenen der CNNs Repräsentationen aus, die denen im menschlichen visuellen System ähnlich sind. Neuronen auf niedrigen Ebenen des Netzwerks haben oft eine Filterantwort, die relativ einfache Merkmale kodiert, vergleichbar z.B. mit Kantendetektoren im frühen visuellen Kortex, während Neuronen auf höheren Ebenen recht komplexe Merkmale kodieren können, z.B. Texturen oder Teile von Gesichtern. Da diese Merkmale relativ generisch sind, ist zu erwarten, dass Transfer auf neuartiges Material gelingt. Es sind heute einige derart vortrainierte Netzwerke verfügbar, die sich mit relativ wenig Aufwand an neues Material anpassen lassen. Der Vergleich der Gewichte für verschiedene Materialtypen erlaubt dann auch Rückschlüsse über deren Unterschiede.

## Fragestellung

Generalisieren die auf Fotos vortrainierten Netzwerke auch auf zeichnerisches Material? Wir berichten von Experimenten, in denen wir das Material des Graphic Narrative Corpus (GNC, Dunst et al., 2017) mit CNNs beschreiben. Der Graphic Narrative Corpus repräsentiert das erste digitale Korpus von englischsprachigen Graphic Novels mit derzeit 130 Titeln. Die ersten Kapitel dieser Werke werden von menschlichen Kodierern annotiert, dabei werden u.a. die Identität und der Ort zentraler Charaktere, Orte von Panels, Sprechblasen und Textboxen (Captions) und Onomatopeia sowie der Text selbst notiert. Außerdem werden Blickbewegungen von Lesern erhoben (Eye-Tracking), um Aufschluss über die Aufmerksamkeitsverteilung auf Seite der Rezipienten zu erhalten.

Die Beschreibung des GNC mit CNNs hat verschiedene Ziele. Erstens erhoffen wir uns Aufschluss über stilistische Unterschiede zwischen Werken und Genres, z.B. mittels Berechnung von Distanzmaßen basierend auf den Modellparametern. Allgemeiner könnte so

der Weg zu einer visuellen Stilometrie aufgezeigt werden, die auch für inhaltliche Bereiche außerhalb der Graphic Novels relevant ist, etwa im Sinne einer computationalen Kunstgeschichte (Saleh & Elgammal, 2015; Manovich, 2015). Zweitens ermöglicht die Beschreibung mit Hilfe der Merkmale tiefer CNNs durch sogenannte Region Proposal Networks (Girshick et al., 2013) die Detektion von Objektklassen. Beispielsweise könnten sich Sprechblasen oder handelnde Charaktere lokalisieren lassen. Wenn Klassen von Objekten automatisiert lokalisiert werden können, erleichtert dies die Arbeit der Annotatoren sehr. Die Ergebnisse können also zurück in das Annotationswerkzeug fließen, um eine Teilautomatisierung zu ermöglichen. Drittens ist aus kognitionspsychologischer Perspektive interessant, welche Merkmale die Aufmerksamkeit auf sich ziehen. Die Korrelation der Netzwerkbeschreibung mit den Blickbewegungsdaten ermöglicht eine Modellierung der Aufmerksamkeitssteuerung auf einem deutlich höheren Auflösungsgrad als die subjektive Beschreibung.

# Methode

Für die Modellierung des Materials nutzen wir die Architektur VGG der Visual Geometry Group in Oxford (Simonyan & Zisserman, 2014), insbesondere VGG-16 und VGG-19. Diese Wahl ist motiviert durch die Einfachheit der Architektur, die die Interpretation der Gewichte erleichtert. Das zugrundeliegende Netzwerk lässt sich aber prinzipiell austauschen; andere Architekturen wie ResNet (He et al., 2015) oder Inception (Szegedy et al., 2015) sind denkbar und sollten ähnlich gute Ergebnisse liefern. Für die Vorhersage der Aufmerksamkeitsverteilung der Leser nutzen wir die Architektur Deep Gaze II (Kümmer et al., 2016). Deep Gaze II ist ein neuronales Netz, das auf VGG-19 aufsetzt und die Antwort einiger dessen Schichten nutzt, um "empirische Salienz" vorherzusagen. Empirische Salienz ist operationalisiert durch Messung von Mauspositionen beim Aufdecken eines verschwommenen Bildes bzw. Messung von Blickbewegungsdaten beim Betrachten von Fotos natürlicher Szenen. Die Fotos sind andere, als die für das Training von VGG-19 benutzten. Man beachte, dass sowohl VGG-19 als auch Deep Gaze II auf Fotos trainiert wurden, also nie Graphic Novels gesehen haben. Da sie jedoch Merkmale und Gewichte herausgebildet haben, die für die Interpretation (von Bildern) der menschlichen Umwelt nützlich sind, kann man vermuten, dass sie sich auch für die Analyse von Zeichnungen eignen. Zwar sind Zeichnungen Abstraktionen, haben aber als solche einen Bezug zur visuellen (Photo-)Realität.

# Ergebnisse

Die Ergebnisse zeigen, dass sich mit Hilfe von Neuronen auf höheren Ebenen der tiefen CNNs recht gut bestimmte Klassen von Objekten lokalisieren lassen. Beispielsweise eignen sich einige Kombinationen von Merkmalen zuverlässig als Sprechblasendetektoren (Abb. 1). Dies ist insofern bemerkenswert, als die Detektion von Sprechblasen sich für klassische Ansätzen der maschinellen Bildbverarbeitung als schwieriges Problem dargestellt hat (Rigaud et al., 2013).

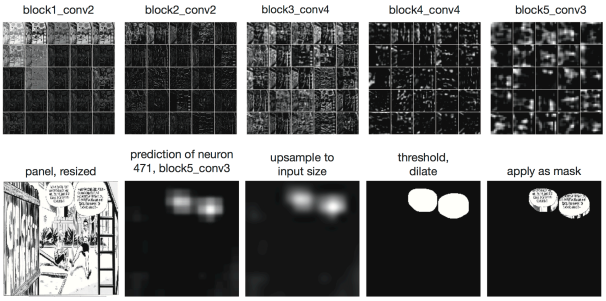


*Abbildung 1. Sprechblasendetektion mithilfe von Convolutional Features.*

Auch für die Erkennung gezeichneter Gesichter eignen sich CNNs, allerdings ist hier ein Training auf Ansichten in verschiedenen Perspektiven (Frontal, Profil) notwendig. Und schließlich lässt sich die empirische Fixationsverteilung mit Deep Gaze II insgesamt sehr überzeugend reproduzieren (Abb. 2). Die CNN-Features kodieren also aufmerksamkeitsrelevante Merkmale.

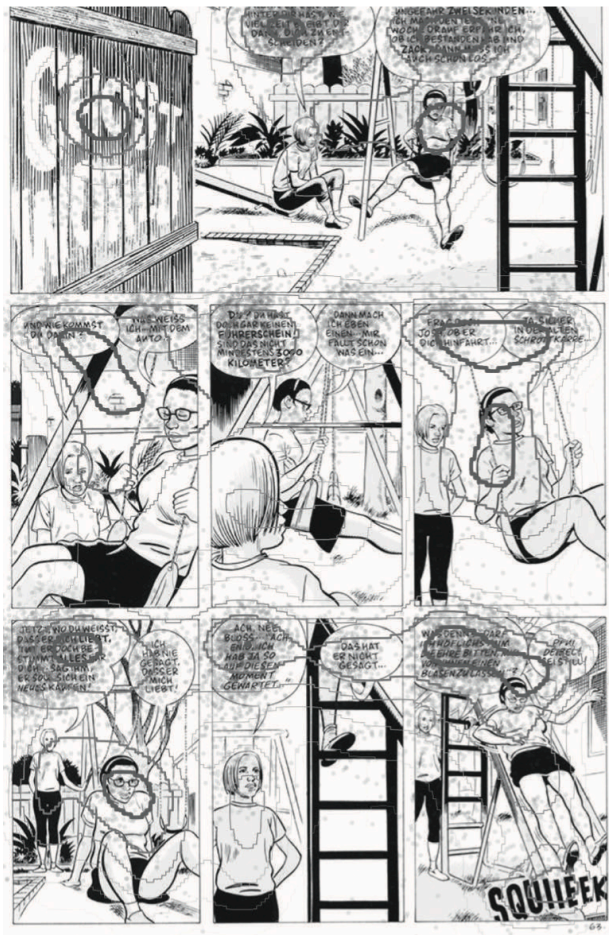


*Abbildung 2. Empirische Fixationsverteilung von 100 Lesern (Punkte) und DeepGaze II-Vorhersagen (Konturlinien).*

Insgesamt eigenen sich auf Fotos trainierte CNNs schon ohne spezifisches weiteres Training recht gut zur Beschreibung gezeichneten Materials in Graphic Novels.

# Diskussion

Die "objektive" Beschreibung eröffnet vielfältige Anwendungen. Einerseits kann, wie oben skizziert, die Annotation visuellen Materials durch Nutzung von vorgeschlagener Regionen deutlich erleichtert werden, etwa vergleichbar mit dem bei verbalen Material durch Verwendung von Optical Character Recognition (OCR) ermöglichten Übergang von kompletter Transkription zum Korrekturlesen. Hier soll angemerkt werden, dass vielversprechende CNN-basierte Ansätze zur Textlokalisation (Sudholt & Fink, 2016) und OCR existieren (Lee & Osindero, 2016). Andererseits sind durch das Vorliegen visueller Merkmale (Features) vielfältige stilometrische Anwendungen denkbar. Zum Beispiel lassen sich aufgrund der Merkmale Ähnlichkeiten verschiedener Zeichner und Künstler berechnen und durch Gruppierung (Clustering) im Merkmalsraum auch Stile definieren. Auch die weitergehende Exploration der Repräsentation auf verschiendenen Schichten des Netzwerks scheint eine vielversprechende Aufgabe weiterer Forschung. Beispielsweise könnte der Vergleich der Antworten auf fotographische versus zeichnerisch abstrahierte Abbilder von Exemplaren einer Kategorie Hinweise auf das Wesen der Abstraktion geben, oder es lassen sich visuelle Merkmale identifzieren, die in besonderem Ausmaß die Aufmerksamkeitszuwendung im Leseprozess und bei der Rezeption von Zeichnungen leiten.

Wir haben beispielhaft aufgezeigt, wie sich Werkzeuge der mathematisch-computationalen Modellierung eignen, um grafisches Material zu analysieren und zu beschreiben. Die Hoffnung ist, dass eine visuelle Stilometrie die Digitalen Geisteswissenschaften im Bereich visuellen Materials in ähnlicher Art und Weise bereichert wie computerlinguistische Ansätze im Bereich der Textanalyse. Digitale Analysen liefern mächtige neue Werkzeuge, die mittel- bis längerfristig auch eine neue Theoriebildung fördern könnten.

# Bibliographie

**Blei, D.** (2012). Probabilistic Topic Models. Communication of the ACM, 55, 77-84.

**Dunst, A., Hartel, R. & Laubrock, J.** (im Druck). The Graphic Narrative Corpus (GNC): Design, Annotation, and Analysis for the Digital Humanities. *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR 2017).*

**Girshick, R., Donahue, J., Darrell, T., & Malik, J.** (2013). Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv*:1311.2524.

**Gumbrecht, H. U.** (2014) **.** Das Denken muss nun auch den Daten folgen. *Frankfurter Allgemeine Zeitung, 12.03.2014, S. 14*, http://www.faz.net/aktuell/feuilleton/geisteswissenschaften/neue-serie-das-digitale-denken-das-denken-muss-nun-auch-den-daten-folgen-12840532.html

**He, K., Zhang, X., Ren, S., & Sun, J.** (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

**Juola, P.** (2006) **.** Authorship Attribution. *Foundations and Trends in Information Retrieval, 1*, 233-334.

**Kümmerer, M., Wallis, T. S. A., & Bethge, M.** (2016) **.** Deepgaze II: Reading fixations from deep features trained on object recognition. *CoRR*, abs/1610.01563.

**Lee, C. & Osindero, S.** (2016). Recursive recurrent nets with attention modeling for OCR in the wild. *CoRR, abs/1603.03101* (CVPR 206).

**Lewandowsky, S. & Farrell, S.** (2011). *Computational Modeling in Cognition: Principles and Practice*. Thousand Oaks: SAGE.

**Manovich, L.** (2015). Data Science and Digital Art History. *International Journal for Digital Art History, 1*, 13-35. http://dx.doi.org/10.11588/dah.2015.1.21631.

**Myung, I. J., & Pitt, M. A.** (2002). Mathematical modeling. In J. Wixted (Ed.), *Stevens' Handbook of*

*Experimental Psychology (Third Edition), Volume IV (Methodology)* (pp. 429–459). New York: John Wiley & Sons.

**Rigaud, C., Burie, J. C., Ogier, J. M., Karatzas, D., & Weijer, J. V. D.** (2013). An active contour model for speech balloon detection in comics. *Proceedings of the 12th International Conference on Document Analysis and Recognition, 1240–1244*.

**Saleh, B. & Elgammal, A. M.** (2015) **.** Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *CoRR*, abs/1505.00855.

**Schich, M., Song, C., Ahn, Y. Y., Mirsky, A., Martino, M., Barabási, A. L., & Helbing, D.** (2014). A network framework of cultural history. *Science, 345(6196)*, 558-562.

**Simonyan, K. & Zisserman, A.** (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

**Sudholt, S. & Fink, G. A**. (2016). Phocnet: A deep convolutional neural network for word spotting in handwritten documents. In 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), pages 277–282.

**Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A.** (2015). Going deeper with convolutions. 2015 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1-9.