KI für Alle 2: Verstehen, Bewerten, Reflektieren

Themenblock Bildklassifikation/Bildsegmentierung:
07_07Transfer_DigitalisierungAlterDokumente

Von Bildern zu Daten. Das digitale Erfassen von alten Dokumenten

Erarbeitet von
Dr. Jacqueline Klusik-Eckert

Projekt ai4all „KI für Alle"
HeiCAD –Heine Center for Artificial Intelligence and Data Science

Lernziele

Nach diesem Video kannst du …

- die Unterschiede bei der Erschließung von medizinischen Bildern und historischen Dokumenten analysieren
- die Unterschiede von OCR und HCR benennen
- die Methoden aus der Bilderkennung aus den medizinischen Beispielen auf das Anwendungsszenario Erschließung alter Dokumente übertragen

Inhalt

Einstieg

Wer etwas über die Geschichte, Kulturgeschichte herausfinden möchte, kommt oft mit googlen nicht weit. Nur ein Bruchteil unserer historischen Dokumente, die Zeitzeugen aus der Geschichte, sind bereits digitalisiert und tiefenerschlossen. Die meisten noch nicht mal wissenschaftlich bearbeitet. In den Archiven weltweit reihen sich kilometerlang Aktenordner und Kladden, die uns Informationen über historische Ereignisse oder Personennetzwerke geben könnten. Das ist so viel Material, dass es Generationen von Wissenschaftler*innen bräuchte, das alles zu sichten, zu kategorisieren, händisch zu transkribieren und dann in Kontexte zu setzten. So, wie man es bisher lange getan hat.

Quelle [1, 2]

Hier können Verfahren der Künstlichen Intelligenz helfen. Doch alte Dokumente und historische Handschriften haben so ihre Tücken.

Quelle [2, 3]

Wofür braucht man KI-gestützte Verfahren für das Erkennen von Texten?

Mittels der Retrodigitalisierung werden aus analogen Objektträgern wie Papier, Pergament oder Karton menschenlesbare Dateiformate. Das heißt aber noch lange nicht, dass wir als Menschen das auch entschlüsseln können, da jemand vor hunderten von Jahren mit Feder und Tusche aufgeschrieben hat.

Quelle [4]

https://www.klassik-stiftung.de/digital/fotothek/digitalisat/40-392/ Heine an Immermann Handschrift https://tcdh01.uni-trier.de/HHP/faksimiles_z/W20B0318_01z.jpg

Schauen wir uns doch mal als Beispiel einen handgeschriebenen Brief von Heinrich Heine an. Kannst du das lesen? Guten Morgen, lieber Immermann! Dann hört es auch bei mir schon auf. Man müsste gelernt haben historische Schriften wie Kurrent oder Sütterlin zu lesen. Und dann müsste man sich noch in das Schriftbild von Heine einsehen.

Quelle [5]

Meine Oma hat mit dem Brief keine Probleme gehabt. Sie hat in der Schule noch Sütterlin schreiben gelernt. Es gibt weltweit über 200 Schriftsysteme durch die Geschichte hindurch und dann kommen noch die, mal gut, mal wirklich schwer lesbaren, meist individuellen, Handschriften dazu.

Quelle: https://www.worldswritingsystems.org/

Das sind viele Muster, die man lernen muss. Eine willkommene Herausforderung für Machine und Deep Learning Verfahren.

Wie funktioniert die Texterkennung?

Bei der Tiefenerschließung von Dokumenten kommt es darauf an, ob man einen gedruckten oder einen handgeschriebenen Text vor sich hat. OCR, das heißt Optical Character Recognition, ist es, wenn man etwas gedrucktes erschließen möchte. Und wenn es um Handschriften geht, nenne man es Handwriting Character Recognition; oder manche sagen auch Handprint Character Recognition.

In einem ersten Schritt werden die Akten erstmal gescannt und als Bildformat abgelegt. Die Metadaten zu dieser Datei können dann neben den physikalischen Angaben auch Informationen über den Inhalt enthalten, insofern das bei der Erfassung auch mit geleistet wird.

Quelle [6]

In einem zweiten Schritt muss man die unterschiedlichen Bildsegmente zuordnen können. Als Menschen können wir binnen Sekunden die Struktur eines Dokuments erfassen: Überschrift, Haupttext, Fußnote, händische Notiz am Rand und siehe da: eine Illustration! Auch wenn man auf den ersten Blick nicht lesen kann, was das gerade bedeuten soll.

Quelle [6]

Doch bei dieser Bilddatei „sieht" in Anführungszeichen der Computer im Moment noch nichts außer Pixel mit Farbwerten. Der Text selbst kann noch nicht erfasst werden. Die unterschiedlichen Bildsegmente müssen also erstmal gefunden werden.

Quelle [7]

Jetzt kommt das Verfahren der optischen Zeichenerkennung zum Zug.
Es zielt darauf ab, gedruckte oder handgeschriebene Texte in maschinenlesbare Texte umzuwandeln. Vortrainierte KI-Verfahren wie neuronale Netzwerke werden dann für die Mustererkennung eingesetzt, um Buchstaben, Zahlen und Symbole in diesem Bild zu identifizieren.

Quelle [7, 8]

Die Verarbeitung erfolgt in mehreren Schritten: Zunächst erfolgt eine Segmentierung, bei der das Bild in Abschnitte für jeden Buchstaben oder jedes Wort unterteilt wird.

Anschließend erfolgt die Merkmalsextraktion, bei der die charakteristischen Eigenschaften der Buchstaben erkannt und analysiert werden. Diese Merkmale werden dann durch die KI-Modelle verwendet, um die Buchstaben zu identifizieren und in einen maschinenlesbaren Text umzuwandeln.

Quelle [9]

Die Genauigkeit von OCR-Systemen hängt stark von der Qualität der Trainingsdaten ab, die für die Anpassung der KI-Modelle verwendet werden. Fortgeschrittene OCR-Systeme verwenden oft Deep-Learning-Modelle wie Convolutional Neural Networks (CNN) oder Recurrent Neural Networks (RNN), um komplexe Muster und Kontextinformationen besser zu erfassen.

Quelle [10]

Zu den bekanntesten Tools und OCR-Modelle gehören Tesseract oder ABBYY FineReader. Für historische Dokumente eignet sich das explizit dafür entwickelte Tool ORC-D.

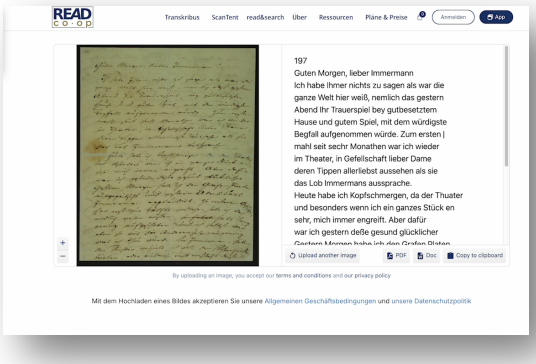
Tools:

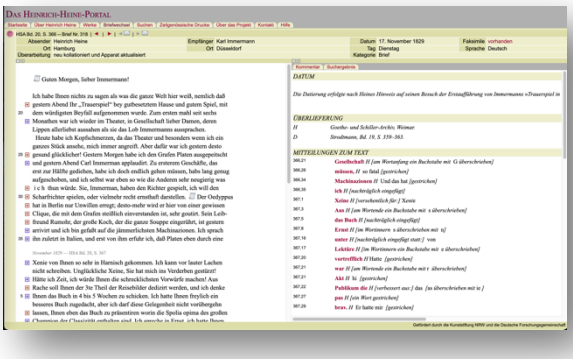
- *Tesseract https://github.com/tesseract-ocr/tesseract, https://en.wikipedia.org/wiki/Tesseract_(software)*
- *OCR-D https://ocr-d.de/de/*
- *ABBYY FineReader*

Das Erkennen von Handschriften geht mittlerweile auch schon mit einigen Smartphones, die ein Erkennungsmodell in der Kamerasoftware eingebaut haben. Das bekannteste,

professionelle Tool für die Entschlüsselung von Handschriften ist aber wahrscheinlich Transkribus. Probieren wir das doch gleich mal mit dem unleserlichen Brief von Heinrich Heine aus.
Transkribus stellt öffentlich eine Anwendung zur Verfügung, die sich dem Modell *The German Giant I* bedient. Ich lade die Bilddatei nun rein und lasse das Netz mal seine Arbeit machen.
Auf der rechten Seite sieht man nun, was als Text alles erkannt wurde. An der linken oberen Ecke wird eine Zahl erkannt. Dann geht es weiter: Guten Morgen, lieber Immermann! Das Ausrufezeichen wird einfach ignoriert.

Quelle [11]





Verglichen mit der digitalen Edition sind noch ein paar Fehler drin. Aber man kann den Inhalt verstehen. Aus einem Bild wird ein Text extrahiert, es wird maschinell suchbar, recherchierbar auf eine neue Ebene zugänglich. Und wenn ich nun weiß, dass ich dieses Dokument nun wirklich brauche, kann ich mich hinsetzten und im Sinne von wissenschaftlichen Editionen mit dem ganzen Regelwerk eine Transkription vornehmen.

Quelle [12]

Oder ich nehme den Text von Transkribus und korrigiere die kleinen Fehler. Wenn ich mehre Dokument der gleichen Schrift oder Handschrift habe, dann lohnt es ich ein Modell dafür zu spezialisieren, sprich: auf die spezifischen Eigenheiten hin trainieren zu lassen.

Unterschied medizinische Bilder und alte Dokumente

Aber warum können wir die vortrainierten Modelle aus der Medizin nicht auch für historische Dokumente verwenden? Schließlich geht es ja auch um das Erkennen von Mustern und Regelmäßigkeiten? Und man hat es mit den gleichen Verfahren der Bildsegmentierung zu tun.
Medizinische Bilder haben in der Regel eine größere Ähnlichkeit. Der Thorax von Menschen sieht beinahe immer gleich aus. Röntgenbilder folgen einer gleichbleibenden Logik.
Bei historischen Dokumenten sieht das ganze schon anders aus. Wir haben nicht nur mit einer Vielzahl von Formaten, Formen, Farben und Kompositionen Layouts zu tun.

Und zu allem Übel kommen noch verschiedene gedruckte Schriften und Handschriften dazu. Da ist es garnicht mal so einfach vortrainierte Modelle weiterzuverwenden.

Quelle [13]

Quellen

Quelle [1] Abbott, Alison. 2017. „The 'Time Machine' Reconstructing Ancient Venice's Social Networks". *Nature* 546 (7658): 341–44. https://doi.org/10.1038/546341a.

Quelle [2] *A Virtual Time Machine for Venice*. 2017. Video. Nature Video. https://www.nature.com/articles/546341a. https://www.youtube.com/watch?v=uQQGgYPRWfs.

Quelle [3] Kaiser, M. (2023). Digitale Sammlungen als offene Daten für die Forschung: Strategische Zielsetzungen der Österreichischen Nationalbibliothek. *Bibliothek Forschung Und Praxis*, *47*(2), 200–212. https://doi.org/10.1515/bfp-2023-0021

Quelle [4] Staatsbibliothek zu Berlin, Reg. 2022. Das Digitalisierungszentrum. https://www.youtube.com/watch?v=0f19mIbLKrk.

Quelle [5] Heine, Heinrich. 2029. „Handschrift Brief von Heinrich Heine an Immermann 4 Seiten vollständig 17.11.1829", 17. November 2029. Handschriften, Reproduktion. Klassik Stiftung Weimar, 40-392

Quelle [6] „Issue 13: OCR". o. J. Europeana PRO. Zugegriffen 5. August 2024. https://pro.europeana.eu/page/issue-13-ocr.

Quelle [7] „GitHub - qurator-spk/eynollah: Document Layout Analysis". o. J. Zugegriffen 5. August 2024. https://github.com/qurator-spk/eynollah?tab=readme-ov-file.

Quelle [8] „OCR-D Workflow Guide - OCR-D". o. J. Zugegriffen 5. August 2024. https://ocr-d.de/en/workflows#workflows.

Quelle [9] „What is OCR4all? | OCR4all". o. J. Zugegriffen 5. August 2024. https://www.ocr4all.org/about/ocr4all.

Quelle [10] Dietrich, Felix. 2021. „OCR vs. HTR or "What Is AI, Actually?"". READ-COOP. Zugegriffen 5. August 2024. https://readcoop.eu/insights/ocr-vs-htr/.

Quelle [11] https://readcoop.eu/de/modelle/the-german-giant-i/

Quelle [12] Digitale Edition: http://www.hhp.uni-trier.de/Projekte/HHP/Projekte/HHP/briefe/01briefevon/adressat/G/index_html?widthgiven=30&letterid=W20B0318&lineref=0&mode=1

Quelle [13] IFLA FAIFE (2020): IFLA Statement on Libraries and Artificial Intelligence. International Federation of Library Associations and Institutions. Verfügbar unter https://repository.ifla.org/handle/123456789/1646.

Disclaimer

Transkript zu dem Video „Bildklassifikation/Bildsegmentierung: Von Bildern zu Daten. Das digitale Erfassen von alten Dokumenten“, Dr. Jacqueline Klusik-Eckert.