Humboldt-Universität zu Berlin
Institut für Informatik
Lehr- und Forschungsgebiet Computer Vision
Prof. Dr. rer. nat. Ralf Reulke

Rudower Chaussee 25
12489 Berlin
Sekr.: +49 30 2093-3901

17. April 2018

Ausschreibung Masterarbeit

Adaptive Texterkennung auf historischen Dokumenten

In vielen Wissenschaftsdisziplinen kommt historischen Dokumenten eine große Bedeutung zu. Oftmals liegen diese jedoch nicht als digitaler Text, sondern allenfalls als digitaler Scan vor. Die Umwandlung dieser Scans in einen digitalen Text ist gerade bei großen Textmengen eine elementare Voraussetzung für die weitere Nutzung.

Herkömmliche Verfahren der Optical Character Recognition (OCR) sind auf die Erkennung maschinengeschriebener Texte spezialisiert, da sie auf einer feststehenden Zeichenmenge operieren und die klare Trennung der Zeichen ausnutzen. In historischen Dokumenten sind diese Einschränkungen durch die Verwendung alter Schriftarten oftmals nicht erfüllt. Dies erfordert die Verwendung neuartiger Ansätze.

In der aktuellsten Forschung haben sich tiefe neuronale Netze (insbesondere Kombinationen von LSTM und CNN) als hilfreich bei der Lösung dieser Problemstellung erwiesen. Hierbei ist wie bei herkömmlichen Ansätzen ein manuelles Training auf einem Teil des zu erkennenden Texts notwendig. Problematisch ist jedoch die Veränderung der Schrift über mehrere Jahre hinweg, wie sie bei historischen Texten oftmals auftritt. Dies würde ein übermäßig aufwändiges Training verlangen. Diese Problematik könnte durch die Verwendung von bspw. Domain-Adversarial Training oder anderen Techniken des unüberwachten Lernens gelöst werden.

Konkret soll ein Datensatz historischer Einwohnerverzeichnisse aus dem 19. Jahrhundert untersucht werden (https://digital.zlb.de/viewer/cms/82/). Hierbei kann auch eine Einbeziehung der Semantik im entstehenden neuronalen Netz erfolgen, um die Erkennung weiter zu verbessern.

Geplant ist eine Zusammenarbeit mit dem Institut für Wirtschaftsgeschichte der HU. Die Betreuung der Arbeit erfolgt durch den Lehrstuhl für Computer Vision.

Voraussetzungen

- Eine der Veranstaltungen Computer Vision, Signalverarbeitung, Mustererkennung oder ähnliche wurde wünschenswerteweise bereits gehört
- Grundlegende Programmiererfahrung, wünschenswerterweise Kenntnisse in Python
- Im Laufe der Arbeit voraussichtlich Einarbeitung in OCR, Verfahren des maschinellen Lernens und TensorFlow notwendig

Referenzen

https://arxiv.org/pdf/1802.10033.pdf
https://arxiv.org/pdf/1505.07818.pdf
(weitere auf Anfrage)

Kontakt

Ansprechpartner	Niklas Deckers	deckersn at hu-berlin.de
	Prof. Ralf Reulke	reulke at informatik.hu-berlin.de
Bearbeitungszeit	6 Monate	
Beginn	Ab sofort	