TECHNISCHE HOCHSCHULE LÜBECK

**Bachelorarbeit**

# Dokumenten und -Texterkennung anhand von Rechnungsbelegen


abgegeben von: Christoph Thomas

Abteilung: Elektrotechnik und Informatik
Studiengang: Informatik/Softwaretechnologie
Erster Gutachter: Prof. Dr. Ehlers



Anfangsdatum: 01.07.2019
Abgabedatum: 31.10.2019

**Arbeitsauftrag**

In der heutigen Software ist Machine Learning zu einem nicht mehr wegzudenkenden Teilgebiet geworden. Durch die Verwendung intelligenter Systeme können Ressourcen wie Zeit und Aufwand gespart werden. Unternehmen, die mit dieser Technologie arbeiten, haben einen ökonimischen Vorteilen gegenüber der Konkurrenz und heben sich von Diesen ab.

Die Automatisierung soll anhand einer künstlichen Intelligenz erfolgen.
Der Erkennungsprozess wird in drei Phasen untergliedert:

1. Mustererkennung

   Ein Klassifikator ist in der Lage Datenmuster in den Rechnungsbelegen zu erkennen und einem Belegtyp zuzuordnen.

2. Belegextraktion

   Auf Basis der Mustererkennung kann eine Extraktion der Informationen durchgeführt werden. Die Positionen der Daten werden durch die Musterkennung ermittelt.

3. Evaluation

   Eine Quantifizierung evaluiert die nachträglich ausgefüllten Felder, fehlerhafte Erkennungen und die durchschnittliche Zeitersparnis. Ergebnisse sollen visuell dargestellt werden.

TECHNISCHE
HOCHSCHULE
LÜBECK

Department of
Electrical Engineering and Computer Science
Dean's Office - Board of Examiners

**Erklärung zur Bachelorarbeit**

Ich versichere, dass ich die Arbeit selbst¨andig, ohne fremde Hilfe verfasst habe.

Bei der Abfassung der Arbeit sind nur die angegebenen Quellen benutzt worden.W¨ortlich oder dem Sinne nach entnommene Stellen sind als solche gekennzeichnet.

Ich bin damit einverstanden, dass meine Arbeit ver¨o↵entlicht wird, insbesondere dass die Arbeit Dritten zur Einsichtnahme vorgelegt oder Kopien der Arbeit zurWeitergabe an Dritte angefertigt werden.

Lübeck, den July 17, 2019

. . . . . .

*(signature)*

**Zusammenfassung der Arbeit / Abstract of Thesis**

| | |
|---|---|
| **Fachbereich:** | Elektrotechnik und Informatik |
| **Studiengang:** | Informatik/Softwaretechnologie B.Sc. |
| **Thema:** | Dokumenten und- Texterkennung von Dokumenten |
| **Zusammenfassung:** | Um die Automatisierung firmeninterner Prozesse zu ermöglichen soll anhand von Machine learning eine Klassifikation von Dokumenten stattfinden. Eine Texterkennung soll letzendlich den Nutzer das Ausfüllen eines Formulars abnehmen. Im Laufe dieser Arbeit sollen Probleme analysiert und geeignete Architekturmuster zur Problemlösung verwendet werden. Anhand von Trainingsdaten wird ein für dieses Problem geeigneter Klassifikator trainiert und im Betrieb verwendet. Letztendlich werden Präzisionsergebnisse, die über eine Schnittstelle gesammelt werden, evaluiert. |
| **Autor** | Christoph Thomas |
| **Betreuender Professor:** | Prof. Dr. Jens Ehlers |
| **WS / SS:** | WS 2019/20 |

# Contents

# 1 Einleitung

## 1.1 Motivation

Bei firmeninternen Schadensabwicklungen werden täglich eine Vielzahl von Bildern von Rechnungsbelegen hochgeladen, die auf einen Archiv-Server gespeichert werden. Darüber hinaus werden Formulare ausgefüllt, um Schäden zu protokollieren. Da die in die Formularfelder eingetragenen Informationen aber in Textform auf den Rechnungsbelegen vorliegen, besteht die Motivation daraus das Ausfüllen des Formulars zu automatisieren.

Machine Learning gewinnt immer mehr an Bedeutung, was sich an diesem Anwendungsfall zeigt. Aus Sicht des Kunden ist die Automatisierung der Schadensabwicklung ein Ersparniss an Zeit und Aufwand. Ökonomisch gesehen erzielt die KI also einen Vorteil für den Kunden. Eine Automatisierung über Bilderkennung würde den Kunden viel Zeit ersparen und somit dem Unternehmen ein enormen Wettbewerbsvorteil verschaffen.

Durch Datenbestände in kann eine künstliche Intelligenz trainiert werden Muster zu erkennen und eigenständig Lösungen für Probleme zu finden. Durch das Füttern der Datenbestände an den Klassifikator wird eine Trainingsphase durchlaufen.
Auf dem Archivserver ist eine Vielzahl von Bildern vorhanden, welche als Trainingsdaten genutzt werden um den Klassifikator auf zukünftige Prognosen vorzubereiten.

Es gibt eine Vielzahl von Klassifikatoren, die jeweils Anwendungsgebiete haben, in denen sie besonders gut Probleme Prognostizieren können. Im Bereich der Bild- und Texterkennung sind Neuronale Netze sehr effektiv.

## 1.2 Problemstellung

Beim Ausfüllen des Formulares geht Zeit und Aufwand verloren. Durch Machine Learning soll dieser Prozess automatisiert werden. Die Adaption bezüglich Machine Learning schafft

gegenüber nicht adaptierenden Unternehmen einen Wettbewerbsvorteil und sichert dem Unternehmen einen hohen Markstellenwert.
Der hochgeladene Beleg soll durch ein Klassifikator kategorisiert und somit der Belegtyp festgestellt werden.
Der bei der Schadensabwicklung hochgeladene Beleg beinhaltet Daten, die in die Formularfelder eingetragen werden müssen. Das Ausfüllen wird durch eine künstliche Intelligenz automatisiert.

## 1.3 Zielsetzung

Durch eine sinnvolle Implementierung zweier Klassifikatoren soll das Ausfüllen eines Schadensabwicklungsformular durch einen Bildupload automatisiert werden. Die Auswahl Der Klassifkationsmodelle sollen für die Aufgaben passend ausgewählt werden. Es soll eine möglichst hohe Präzisionsrate für Probleme erzielt werden. Weiterhin soll eine Schnittstelle die Ergebnisse der Klassifikation abfangen und auf einem Dashboard sammeln um diese dann zu visualisieren.
Es soll eine Stufenweise Klassifikation stattfinden:

1. Klassifikation der Belege
2. Extraktion der Informationen auf Basis der Belegart durch OCR (optical character recognition)

Die Trainingsphase soll den Klassifikator effektiv auf bevorstehende Mustererkennungen vorbereiten. Die vorhandenen Trainingsdaten sollen aufbereitet werden und dem Klassifikationsmodell übergeben werden.

## 1.4 Gliederung der Arbeit

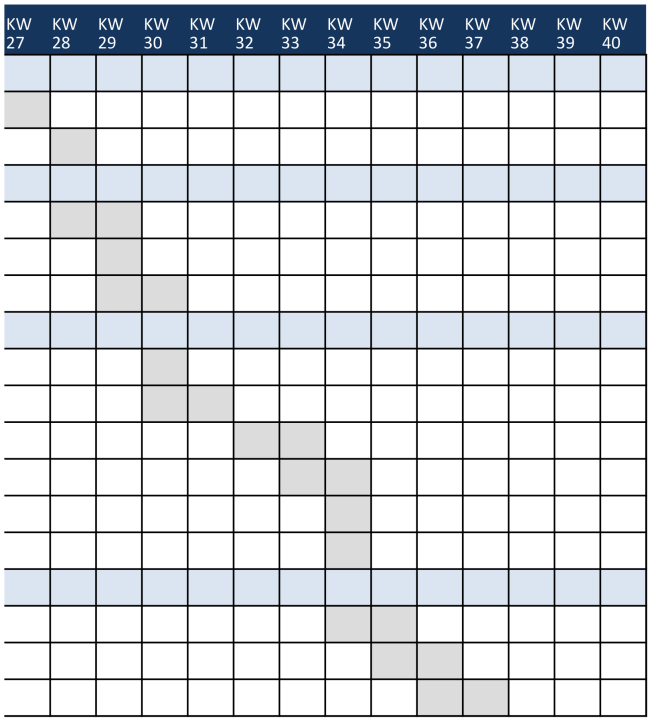
Die Arbeit unterteilt sich in momentan 3 Kernbereiche:

- Dokumentenerkennung (Document Image Analysis)
- Bilderkennung mit einem neuronalen Netz
- OCR mit Tesseract

## 1.5 Meilensteinplan

| | |
|---|---|
| **Startdatum** | 01.07.19 |
| **Enddatum** | 01.10.19 |
| **Fortschritt** | 5,00% |

| Aufgaben | Start | Ende | Tage | Status | |
|---|---|---|---|---|---|
| **Vorarbeiten** | | | | | |
| Aufgabenanalyse | 1.7 | 5.7 | 4 | 75,00% | |
| Informationsbeschaffung | 5.7 | 10.7 | 5 | 25,00% | |
| **Modellierung** | | | | | |
| Klassifikationsmodell | 10.7 | 15.7 | 5 | 25,00% | |
| Architektur | 15.7 | 20.7 | 5 | Nicht begonnen | |
| Trainingsumgebung | 20.7 | 22.7 | 2 | Nicht begonnen | |
| **Entwicklung** | | | | | |
| Abhängigkeiten, Umgebung | 22.7 | 24.7 | 2 | Nicht begonnen | |
| Architektur umsetzen | 24.7 | 4.8 | 11 | Nicht begonnen | |
| Trainingsphase | 4.8 | 12.8 | 8 | Nicht begonnen | |
| Evaluation der Ergebnisse | 12.8 | 19.8 | 7 | Nicht begonnen | |
| Optimierung | 19.8 | 22.8 | 3 | Nicht begonnen | |
| Kommentierung | 22.8 | 23.8 | 1 | Nicht begonnen | |
| **Evaluation** | | | | | |
| Pipeline definieren | 23.8 | 29.8 | 6 | Nicht begonnen | |
| Ergebnisse visualisieren | 29.8 | 6.9 | 8 | Nicht begonnen | |
| Auswertung | 6.9 | 11.9 | 5 | Nicht begonnen | |



# 2 Anforderungsanalyse

Dieser Chapter fügt sich in den Bereich der Anforderungsanalyse

## 2.1 Anforderungen

## 2.2 Probleme

Durch die stetige Weiterentwicklung von DIA ist die Vielfalt der Probleme gewachsen[Ten19]:

- Binarisierung - Klassifizieren welche Pixel im Vordergrund un welche im Hintergrund ist
- Korrektur - Korrigieren der perspektivischen Verzerrung. Dies passiert wenn die Kamera nicht orthogonal zum Physikalischen Dokument positioniert ist
- Seitensegmentierung - Unterteilen einer Seite in homogene Komponenten wie Text, Bilder und Grafiken
- Textsegmentierung - Unterteilen eines Textabschnitts in Zeilen
- Dokumentklassifizierung - Klassifizierung des Dokumenttyps
- Schrift und - Spracherkennung - Klassifizieren mit welcher Schrift geschrieben wurde und welche Sprache benutzt wird
- Zeichensegmentierung - Segmentiert ein Wortbild in individuelle Charaktere
- Grafische Erkennung - Erkennung von graphischen Komponenten
- Tabellenlayout-Struktur Erkennung - Wiederherstellen der Zeilen und- Spaltenstruktur von einem Bild einer Tabelle
- Öffentlich genutzte Text und Grafiken erkennen - In der Öffentlichkeit genutzen Text und Grapfiken klassifizieren, wie beispielsweise Nummerschilder und Straßenzeichen
- Identifikation des Verfassers - Das bestimmen von indivuellen Charakteristiken des Verfassers des Textes, wie beispielsweise Geschlecht oder Alter
- Messung der Dokumentenqualität

- Handgeschriebenen Text erkennen - Texterkennung bei handgeschrieben Text von individuellen Verfassern

## 2.3 Vorgehensweise in DIA

Die typische Bildverarbeitungspipeline für Dokumente besteht aus drei allgemeinen Schritten.

1. Vorverarbeitung - Das preprocessing der Dokumente beinhaltet das Entfernen von Rauschen und Unschärfe, Korrektur, Entzerrung und Binarisierung.

2. Layout-Analyse - Verstehen der Dokumentstruktur um Regionen die im Interesse (Regions of Interest) liegen zu identifizieren.

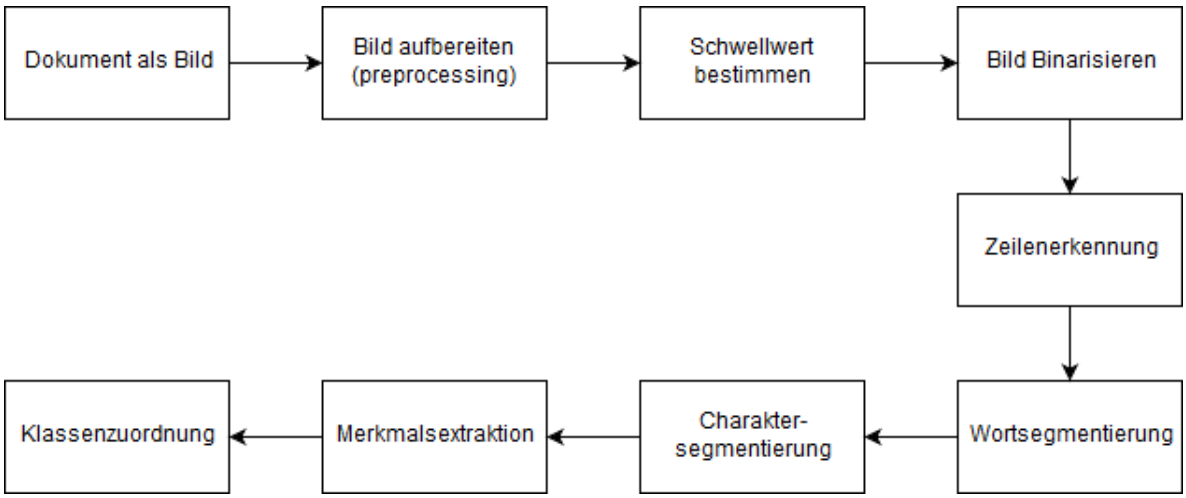
3. Erkennung - Extrahieren von den Informationen aus jeder Rol.

## 2.4 Optical Charakter Recognition (OCR)

Eine Technik von DIA ist die optische Zeichenerkennung (OCR), die auf das Erkennen gedruckter Zeichen ausgelegt ist. Die Technologie findet derzeit bei der automatisierten Weiterleitung von E-Mails statt, indem Postleitzahlen von Briefumschlägen geparst werden. Generell kann OCR als möglicherweise gelöstest Problem für maschinell gedruckte und gescannte Bürodokumente in Englischer Schrift betrachtet werden. Es werden Genauigkeiten von mehr als 99% erzielt. Jedoch gibt es über 3000 geschriebene Sprachen und bei der optischen Zeichenerkennung ist bei der Mehrheit der Sprachen ein ungelöstes Problem, welches sich in laufender Forschung befindet.

## 2.5 Funktionsweise von OCR

In 3.5 ist die seqentuelle Work ow von dem Erkennen eines Textes abgebildet. Das Bild wird als Dokument hochgeladen und mit preprocessing wird das Bild fur den Klassierungsprozess aufbereitet. Abhangig von den bestimmten Schwellwert wird das Bild binarisiert. Texte werden als weie Pixel dargestellt, wahrend alle darumliegende Pixel zu schwarz konvertiert werden. Somit hebt sich der Text komplett vom Hintergrund ab. Nun konnen Zeilenweise die Worter segmentiert werden, um diese wiederum in Charakter zu segmentieren. Die Charaktere werden als einzelnde als Merkmale extrahiert und klassiziert.


Dokument als Bild
Bild aufbereiten (preprocessing)
Schwellwert bestimmen
Bild Binarisieren
Zeilenerkennung
Wortsegmentierung
Charakter-segmentierung
Merkmalsextraktion
Klassenzuordnung

# 3 Dokumentenerkennung

Dieser Chapter fügt sich in den Bereich der Document Image Analysis (DIA).

## 3.1 Forschungsstand

DIA bezeichnet allgemein eine große Gruppe von Techniken, die visuelle Informationen charakterisieren können. Dia gehört zu den ältesten Bereichen der Informatik. 1913 Erfand Edmund Fournier das Optophon, mit dem dunkle Schriftzeichen auf einem Stück Papier erkannt werden können und als Töne interpretiert wurden.
Heutzutage durchlaufen moderne DIA-Algorithmen Rasterbilder, die mit Scannern, Kameras oder anderen Digitalisierungsgeräten erzeugt wurden.

## 3.2 Probleme

Durch die stetige Weiterentwicklung von DIA ist die Vielfalt der Probleme gewachsen[Ten19]:

- Binarisierung - Klassifizieren welche Pixel im Vordergrund un welche im Hintergrund ist
- Korrektur - Korrigieren der perspektivischen Verzerrung. Dies passiert wenn die Kamera nicht orthogonal zum Physikalischen Dokument positioniert ist
- Seitensegmentierung - Unterteilen einer Seite in homogene Komponenten wie Text, Bilder und Grafiken
- Textsegmentierung - Unterteilen eines Textabschnitts in Zeilen
- Dokumentklassifizierung - Klassifizierung des Dokumenttyps
- Schrift und - Spracherkennung - Klassifizieren mit welcher Schrift geschrieben wurde und welche Sprache benutzt wird
- Zeichensegmentierung - Segmentiert ein Wortbild in individuelle Charaktere
- Grafische Erkennung - Erkennung von graphischen Komponenten
- Tabellenlayout-Struktur Erkennung - Wiederherstellen der Zeilen und- Spaltenstruktur von einem Bild einer Tabelle

- Öffentlich genutzte Text und Grafiken erkennen - In der Öffentlichkeit genutzen Text und Grapfiken klassifizieren, wie beispielsweise Nummerschilder und Straßenzeichen
- Identifikation des Verfassers - Das bestimmen von indivuellen Charakteristiken des Verfassers des Textes, wie beispielsweise Geschlecht oder Alter
- Messung der Dokumentenqualität
- Handgeschriebenen Text erkennen - Texterkennung bei handgeschrieben Text von individuellen Verfassern

## 3.3 Vorgehensweise in DIA

Die typische Bildverarbeitungspipeline für Dokumente besteht aus drei allgemeinen Schritten.

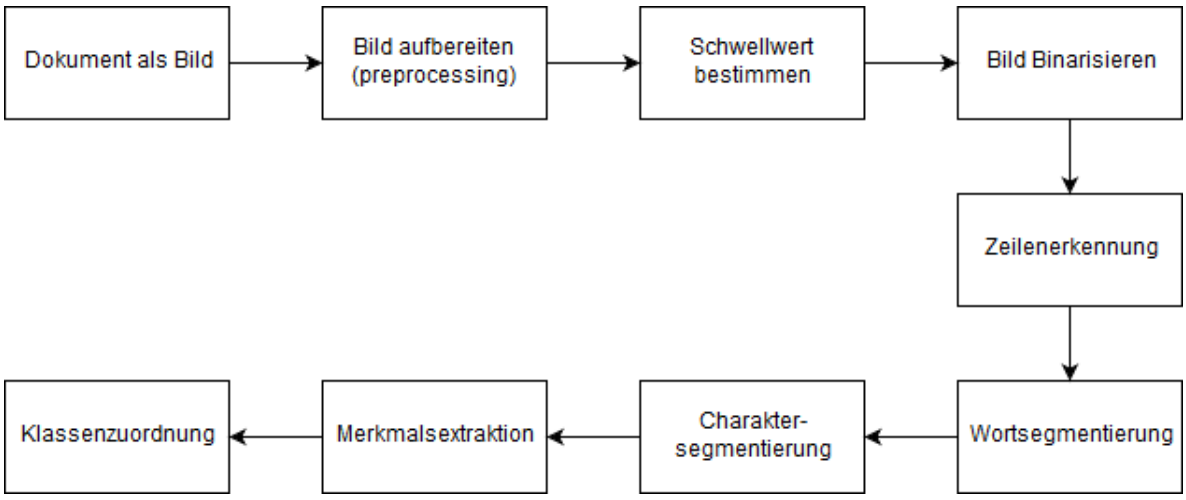
1. Vorverarbeitung - Das preprocessing der Dokumente beinhaltet das Entfernen von Rauschen und Unschärfe, Korrektur, Entzerrung und Binarisierung.
2. Layout-Analyse - Verstehen der Dokumentstruktur um Regionen die im Interesse (Regions of Interest) liegen zu identifizieren.
3. Erkennung - Extrahieren von den Informationen aus jeder Rol.

## 3.4 Optical Charakter Recognition (OCR)

Eine Technik von DIA ist die optische Zeichenerkennung (OCR), die auf das Erkennen gedruckter Zeichen ausgelegt ist. Die Technologie findet derzeit bei der automatisierten Weiterleitung von E-Mails statt, indem Postleitzahlen von Briefumschlägen geparst werden. Generell kann OCR als möglicherweise gelöstest Problem für maschinell gedruckte und gescannte Bürodokumente in Englischer Schrift betrachtet werden. Es werden Genauigkeiten von mehr als 99% erzielt. Jedoch gibt es über 3000 geschriebene Sprachen und bei der optischen Zeichenerkennung ist bei der Mehrheit der Sprachen ein ungelöstes Problem, welches sich in laufender Forschung befindet.

## 3.5 Funktionsweise von OCR

In 3.5 ist die seqentuelle Work ow von dem Erkennen eines Textes abgebildet. Das Bild wird als Dokument hochgeladen und mit preprocessing wird das Bild fur den Klassierungsprozess aufbereitet. Abhangig von den bestimmten Schwellwert wird das Bild binarisiert. Texte werden als weie Pixel dargestellt, wahrend alle darumliegende Pixel zu schwarz konvertiert




werden. Somit hebt sich der Text komplett vom Hintergrund ab. Nun konnen Zeilenweise die Worter segmentiert werden, um diese wiederum in Charakter zu segmentieren. Die Charaktere werden als einzelnde als Merkmale extrahiert und klassiziert.

# List of Figures

# List of Tables

# Bibliography

[Ten19] Christopher Alan Tensmeyer. *Deep Learning for Document Image Analysis*. https://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=8389&context=etd. Apr. 2019.