SEMANTISCHE SEGMENTIERUNG MIT HILFE NEURONALER NETZWERKE ZUR EFFIZIENTEN VERARBEITUNG DIGITALISIERTER DOKUMENTE

Prof. Dr. Klaus Jung[a]

[a] Fachbereich Informatik, Kommunikation und Wirtschaft, Hochschule für Technik und Wirtschaft Berlin, Deutschland, klaus.jung@htw-berlin.de

KURZDARSTELLUNG: Bei der Digitalisierung von Dokumenten lassen sich hochqualitative PDF Dateien mit sehr kleiner Dateigröße erzeugen, sobald Informationen über den Inhaltstyp von Seitenbereichen verwendet werden. Eine solche Segmentierung der Seitenbilder wird mit Hilfe von Techniken des maschinellen Lernens vorgestellt. Dabei kommen Fully Convolutional Networks zum Einsatz. Neben einem Vergleich der Ergebnisse mit konventionellen Ansätzen wird auf das Erzeugen geeigneter Trainingsdaten eingegangen. Anwendungen zur Klassifizierung und automatischen Verschlagwortung von Inhalten sind möglich.

1. EINFÜHRUNG

Bei der digitalen Konversion der Bestände von Museen, Bibliotheken und Archiven steht die authentische, hochqualitative Reproduktion der Originale im Vordergrund. Im Hinblick auf die Bereitstellung der digitalisierten Inhalte für eine größere Öffentlichkeit stellt sich aber auch die Frage nach effektiven digitalen Formaten, die diese hohe Qualität mit vertretbarem Datenvolumen abbilden können. Im Bereich der Dokumentenarchivierung hat sich der PDF/A-Standard [1] etabliert, der jedoch sehr viel Spielraum für das Einbetten der Seitenbilder zulässt. Eine gängige Technik ist die Mixed Raster Content Verarbeitung [2], bei der eine eingescannte Seite in Bereiche verschiedener Inhaltstypen (z.B. Text oder Bild) klassifiziert wird (Abb. 1). Je nach Inhaltstyp werden Rastergrafiken mit unterschiedlichen Verfahren (z.B. FAX G4, JBIG2, JPEG, JPEG2000) in das PDF eingebunden. Zentraler Punkt dieses Ansatzes ist das Auffinden solcher Bereiche, d.h. die semantische Segmentierung des Seitenbildes.

Existierende Techniken zur Segmentierung von eingescannten Dokumenten versuchen oft über Domänenwissen ein gutes Resultat zu erzielen [3]. Es wird also das „Wissen" darüber, was einen fotorealistischen Bildanteil von einem Text unterscheidet, fest implementiert. Was zunächst einfach klingt, ist es in der Praxis aber nicht. Die Grenzen zwischen Bild und Text sind nicht eindeutig. Bilder können Textanteile enthalten, selbst schwarzer Text enthält im Scan farbige Pixel. Noch komplizierter wird es, sobald die Anzahl zu identifizierender Inhaltstypen erhöht wird: Druckschrift, Handschrift, fotorealistisches Bild, gemaltes Bild, Strichzeichnung, etc.

Aufgrund der wachsenden Rechenleitung insbesondere auf Grafikkarten werden künstliche neuronale Netzwerke derzeit in vielen Anwendungsbereichen eingesetzt. Im Bereich der Objekterkennung erzielen Convolutional Neural Networks zunehmend bessere Ergebnisse [4]. Hierbei ist kein „Vorwissen" über die Inhalte mehr notwendig. Die Systeme lernen die notwendigen Kenntnisse selbständig. Voraussetzung dafür sind sehr viele Trainingsdaten. Neuste Arbeiten verwenden die Variante der Fully Convolutional Networks zur pixelgenauen Segmentierung von Bildinhalten [5]. Die vorliegende Arbeit verknüpft den Bereich der Dokumentenverarbeitung mit den auf neuronalen Netzwerken basierenden Ansätzen. Fully Convolutional Networks werden zur Segmentierung eingescannter Dokumentenseiten verwendet und mit den Resultaten konventioneller Ansätze [3] verglichen. Neben dem Ziel des Erzeugens hochoptimierter PDF-Dateien zum Bereitstellen digitalisierter Inhalte kann diese Technik im Ausblick auch zur Klassifizierung und automatischen Verschlagwortung von Inhalten genutzt werden.

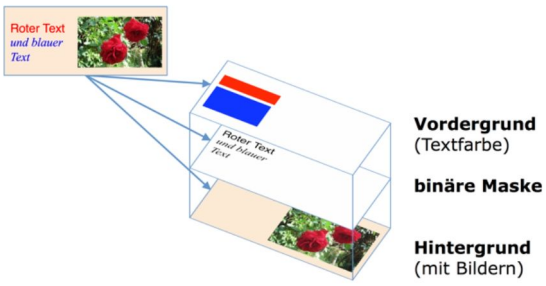


Abb. 1: Schema der Mixed Raster Content (MRC) Verarbeitung

2. MRC KOMPRESSION

Werden Dokumente durch Abfotografieren oder Einscannen digitalisiert, so beträgt die Rohdatenmenge für eine DIN-A4 Seite in 300 dpi ca. 26 MB. Für die Archivierung eines „digitalen Masters“ werden die Rastergrafiken meist nur sehr moderat komprimiert. Für die Bereitstellung z.B. als Download empfiehlt sich eine Verarbeitung, bei der eine noch gute visuelle Qualität mit deutlich besseren Kompressionsraten erzielt wird. Abbildung 2 zeigt vergrößerte Ausschnitte einer Verarbeitung mit unterschiedlichen Ansätzen. Bitonale Verfahren wie FAX G4 oder JBIG2 sind für Text sehr gut geeignet, können aber keine Farbe wiedergeben (Abb. 2(b)). Verfahren für fotorealistische Bildanteile wie JPEG oder JPEG2000 führen zu deutlichen Artefakten im Textanteil (Abb. 2(c)). Erst die Kombination von drei Layern, wie in Abbildung 1 schematisch dargestellt, führt zu keinen nennenswerten Artefakten bei Text- und Bildanteilen (Abb. 2(d) mit JPEG2000 für den Vorder- und Hintergrundlayer). Zum besseren Erkennen der Artefakte empfiehlt es sich, in die elektronische Version dieses Artikels tiefer hinein zu zoomen.

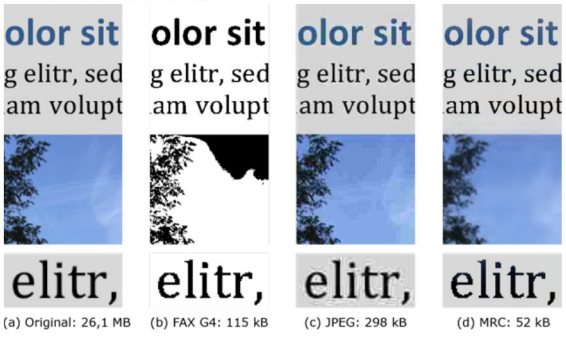


Abb. 2: Vergleich von Kompressionsverfahren

In der Praxis werden die nicht sichtbaren, ausmaskierten Bereiche von Vorder- und Hintergrundlayer durch einen Farbverlauf so angepasst, dass harte Kanten, die zu Artefakten führen würden, weitestgehend vermieden werden. Insbesondere werden im Hintergrundlayer die durch das Entfernen des Textes entstandenen Löcher mit der Umgebungsfarbe gefüllt (Abb. 3). Für die Rekonstruktion entscheidet dann die binäre Maske, ob die Farbe des Pixels aus dem Vordergrund- oder dem Hintergrundbild genommen wird. Eine derartige Verknüpfung von Layern lässt sich leicht mit dem PDF Format realisieren. Zu Archivierungszwecken sollte dabei die standardisierte PDF/A-Variante benutzt werden [1]. Damit wird unter anderem sichergestellt, dass die Farben der Layer mit entsprechenden Farbprofilen genau spezifiziert sind.

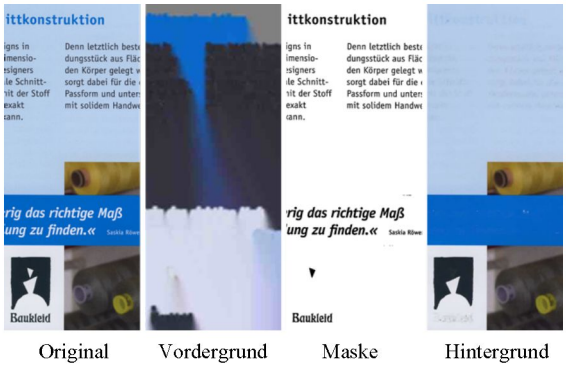


Abb. 3: MRC Layer

Vorder- und Hintergrundlayer werden oftmals mit reduzierter Auflösung in das PDF eingebettet. Damit lässt sich die Dateigröße weiter reduzieren. Lediglich die Maske bleibt in voller Auflösung erhalten, um die Lesbarkeit des Textes nicht zu beeinträchtigen. Ob eine solche Auflösungsreduktion vorgenommen wird, hängt stark von der Anwendung ab. Geht es vor allem um die Lesbarkeit des Textes, mag dies akzeptabel sein. Sind die Bildanteile wichtig, sollte die Auflösung des Hintergrundlayers nicht oder nur wenig reduziert werden.

Die MRC Verarbeitung steht und fällt mit der Güte der Segmentierung. Durch die binäre Maske bleibt der Text zwar gut lesbar, wirkt aber im Vergleich zum Original leicht ausgestanzt. Wird in fotorealistischen Bildanteilen Text erkannt, so entstehen Artefakte wie sie Abbildung 4 in einer Vergrößerung wiedergibt.

Original MRC

Abb. 4: MRC Rekonstruktion mit Artefakten aufgrund fehlerhafter Segmentierung

3. NEURONALE NETZWERKE

Algorithmen, die versuchen, das menschliche Gehirn nachzubilden, sind seit den 60er Jahren bekannt. In den 90er Jahren ließ das Interesse daran jedoch stark nach. Mit der damaligen Hardware konnten im Vergleich zu anderen Ansätzen keine besseren Ergebnisse erzielen werden. Aufgrund leistungsstarker GPUs und der Möglichkeit, solche Netzwerke mit sehr vielen Beispieldaten zu trainieren, hat sich die Situation vor einigen Jahren grundlegend verändert. Im Bereich der Bildklassifikation erzielen Netzwerke wie AlexNet [4], VGG [6] oder GoogLeNet [7] Ergebnisse, die zuvor nicht erreicht werden konnten. Aktuell zählen daher neuronale Netzwerke zu den besten Techniken im Bereich des maschinellen Lernens. Ihr großer Vorteil liegt in der Universalität dessen, was sie erlernen können. Ihr Nachteil in der Tatsache, dass sehr viele Trainingsdaten benötigt werden, um ein gut trainiertes Netzwerk zu erzeugen.

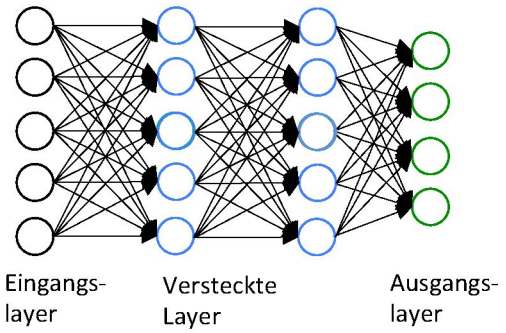


Abb. 5: Neuronales Netzwerk

Abbildung 5 veranschaulicht das Grundprinzip eines neuronalen Netzwerks bestehend aus miteinander verbundenen Neuronen. Für die Klassifikation werden am Eingangslayer die Farbwerte der einzelnen Pixel angelegt. Diese bestimmen die Aktivität der dortigen Neuronen. Jede Verbindung besitzt eine Gewichtung, mit der die Weiterleitung der Aktivität an Neuronen der tieferen Schichten beeinflusst wird. Die Neuronen des Ausgangslayers stellen die zu detektierenden Klassen dar. Im Falle der Erkennung von Text und Bild wären dies zwei Neuronen. Nach geeigneter Normierung kann die Aktivität dieser Neuronen als Wahrscheinlichkeit für die Erkennung der entsprechenden Klasse interpretiert werden. Das Neuron mit der höchsten Aktivität bestimmt die detektierte Klasse.

Für das Training werden die Gewichte des Netzwerks zufällig initialisiert. Am Eingangslayer werden nacheinander Trainingsbilder angelegt. Die berechnete Aktivität am Ausgang wird mit der bekannten Klasse des Trainingsbildes verglichen. Die Abweichung wird benutzt, um im Lernvorgang die Gewichte der Verbindungen anteilig anzupassen. Diese Prozedur wird mit allen Trainingsbildern mehrmals (in Epochen) wiederholt.

In der vorliegenden Arbeit kommen neuronale Netzwerke in Form von Convolutional Neural Networks beim Training und Fully Convolutional Networks beim Segmentieren von Seiteninhalten zum Einsatz.

3.1 CONVOLUTIONAL NEURAL NETWORKS

Im Gegensatz zum in Abbildung 5 dargestellten Grundprinzip sind bei Convolutional Neural Networks die Neuronen nicht mit allen Neuronen der vorherigen Schicht verbunden. In Analogie zum Kernel eines linearen Filters kommen eine Anzahl von Kopien (z.B. 96 Features) eines Kernels begrenzter Größe (z.B. 5x5) zum Einsatz, deren Werte den Gewichten entsprechen. Die Weiterleitung der Aktivität von Neuronen geschieht durch eine Faltungsoperation des jeweiligen Kernels mit den Daten der vorherigen Schicht. Ohne hier auf Fragen der Randbehandlung einzugehen, ergäbe sich für den ersten versteckten Layer eine Anzahl von Neuronen, die der Größe des Eingangsbildes multipliziert mit der Anzahl der Features entspricht, also ein dreidimensionaler Datensatz Bildbreite × Bildhöhe × Anzahl der Features. Beim Eingangslayer beträgt die Anzahl der Features drei, was der Anzahl der Farbkanäle entspricht. Um die Daten zu aggregieren, wenden einige Schichten

zusätzlich eine Unterabtastung an, was die Breite und Höhe des Datenarrays entsprechend reduziert. Dies geschieht so lange, bis am Ende die ersten beiden Dimensionen 1x1 sind. In einer letzten sogenannten vollvernetzten Schicht werden die verbliebenen Neuronen mit einem Ausgangslayer verbunden, dessen Neuronenanzahl der Zahl zu erkennender Klassen entspricht. Abbildung 6 zeigt eine vereinfachte schematische Darstellung des in dieser Arbeit verwendeten Netzwerks.

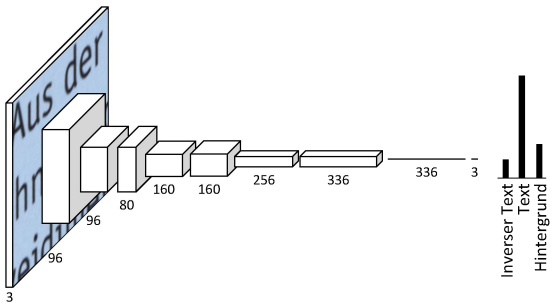


Abb. 6: *Schema eines Convolutional Neural Network*

3.2 FULLY CONVOLUTIONAL NETWORKS

Das in vorherigen Abschnitt vorgestellte Netzwerk verarbeitet nur sehr kleine Bilder, um eine Vorhersage über die Bildklasse zu treffen. Anwendungen im Bereich der Objekterkennung skalieren daher die zu analysierenden Bilder stark herunter. Die in [4], [6] und [7] vorgestellten Netzwerke verarbeiten typischer Weise Bilder der Größe 224x224 Pixel. Um eine 2500x3200 Pixel große DIN-A4 Seite zu verarbeiten, müsste für jeden der acht Millionen Pixel eine kleine Region ausgeschnitten und dem Netzwerk zugeführt werden. Das würde auch mit high-end Grafikkarten zu inakzeptablen Rechenzeiten führen. Für die hier vorgestellte semantische Segmentierung werden die in [5] beschriebenen Techniken adaptiert. Die Grundidee besteht darin, dass die letzte vollvernetzte Schicht durch eine Faltung mit einem 1x1 Kernel ersetzt wird. Dadurch liefert das Netzwerk beim Anlegen eines größeren Bildes nicht nur ein Mal Wahrscheinlichkeiten für jede trainierte Klasse, sondern ein ganzes Array solcher Wahrscheinlichkeiten. Ein Element dieses Arrays steht dann für die erkannte Klasse eines Teilbereichs des Eingangsbildes. Das Netzwerk aus Abb. 6 nimmt vier Unterabtastungen um den Faktor 2 vor. Ein Element des Ausgangsarrays steht in diesem Fall für einen Block von 16x16 Pixeln. Aus einer Aussage über ein einziges Bild, wird eine Heatmap über die räumliche Verteilung der detektierten Klassen. Abbildung 7 illustriert, wie aus dem Convolutional Neural Network aus Abb. 6 ein Fully Convolutional Network wird.

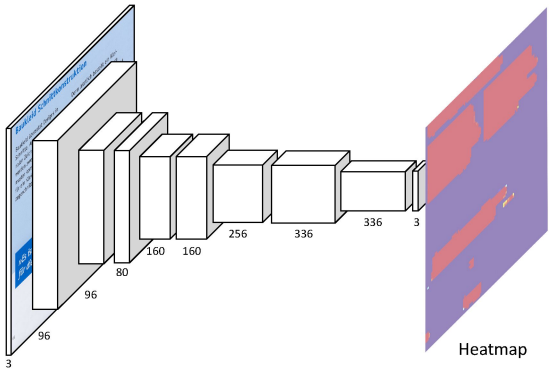


Abb. 7: *Schema eines Fully Convolutional Network*

In [5] werden weitere Maßnahmen beschreiben, um die Segmentierung von einer blockweisen Genauigkeit auf einzelne Pixel zu erhöhen. Für das hier beschriebene Verfahren ist eine derart hohe Genauigkeit nicht notwendig. Text- und Bildbereiche in Dokumenten folgen selten einer so feinen Grenze. Die Techniken zur pixelgenauen Segmentierung kommen daher nicht zur Anwendung.

4. NETZWERKDESIGN

Das in dieser Arbeit verwendete Netzwerk ist eine starke Vereinfachung des GoogLeNet [7]. Die Vereinfachungen wurden in Voruntersuchungen schrittweise durchgeführt, um ein gutes Verhältnis aus Bearbeitungszeit und Erkennungsgenauigkeit zu erzielen. Wird die typische Größe eingescannter Buchstaben in 300 dpi betrachtet, so erweist sich ein Eingangsbild mit 112x112 Pixeln als ausreichend, um die lokalen Eigenschaften des Dokuments gut zu erkennen. Größere Bilder beeinflussen die Erkennung negativ durch irrelevante Strukturen in Randbereich. Zu kleine Bilder enthalten zu wenig Kontext, um Text verlässlich detektieren zu können.

Eine weitere Vereinfachung besteht in der Komplexität des Netzes selbst. Da nur sehr wenige Klassen zu erkennen sind, kann die Kapazität des Netzes reduziert werden. Dies betrifft im Wesentlichen die Tiefe des Netzwerks, aber auch die Anzahl der parallelen Stränge bei den Abschnitten mit Inception [7].

GoogLeNet enthält neun solcher Abschnitte mit jeweils vier Strängen. Das hier verwendete Netz reduziert dies auf fünf Abschnitte mit nur zwei Strängen. Zusätzlich wird die Anzahl der Features in den Konvolutionsschichten verringert.

Zum Training wird ein Convolutional Neural Network verwendet, während bei der Erkennung die Fully Convolutional Variante zum Einsatz kommt (Abb. 6 und 7). Ohne zusätzliche Maßnahmen wäre die Randbehandlung in beiden Varianten verschieden. Im Innern des vollen Netzwerks ist keine Randbehandlung notwendig, während die Faltungen beim Trainieren der kleinen Eingangsbilder eine Randbehandlung vornehmen müssen. Daher wird das Trainingsnetzwerk so entworfen, dass es eingangsseitig einen vergrößerten Bildbereich verarbeitet, der nach der letzten Faltung noch 7x7 Werte liefert. Davon wird jedoch nur die zentrale 1x1 Position weiterverarbeitet.

5. TRAINING

Zum Training werden 72 Dokumentenseiten in 300 dpi eingescannt. Diese stammen aus zwei Dokumenten mit unterschiedlichem Charakter. Das erste Dokument ist ein Katalog eines Wissenschaftsverlags. Es enthält viel Text und Bildanteile, die sehr sauber vom Text getrennt sind. Das zweite Dokument ist eine Zeitschrift, bei der Text und Bild stark ineinander verwoben sind. Oft wird hier Text über ein Bild gesetzt. Text umfließt Bilder nicht nur in Rechtecken.

Die Trainingsdaten werden manuell annotiert. Dazu werden Textzeilen bzw. einzelne Worte mit einem Rechteck markiert. Es wird zwischen Text und inversem Text unterschieden. Bei inversem Text sind die Buchstaben heller als der Hintergrund. Diese Unterscheidung ist für die nachfolgende MRC Verarbeitung wichtig. Es ist darauf zu achten, dass nach einer Binarisierung der Textbereiche die Buchstaben und nicht deren Hintergrund in die Maske gelangen. Abbildung 8 zeigt einen Ausschnitt aus einer annotierten Seite.

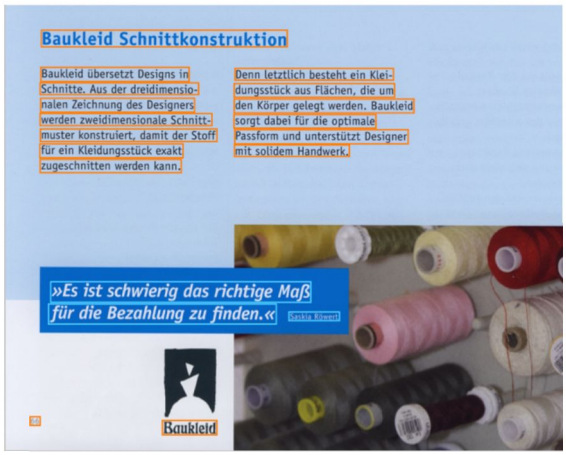


Abb. 8: Zum Training annotierte Textseite

Da sich die Grenzen zwischen Text und Hintergrund nicht pixelgenau definieren lassen, wird aus der Annotation eine Ground Truth berechnet, die an den Übergängen Pixel einer undefinierten Klasse enthält. Diese sind in Abbildung 9 weiß dargestellt. Aus den definierten Bereichen werden Trainingsbilder der Größe 112x112 an zufälligen Positionen entnommen. Dabei wird nur darauf geachtet, dass das Zentrum des Bereichs in einer definierten Klasse liegt. In Abbildung 10 sind einige dieser Bilder für die drei betrachteten Klassen abgebildet.

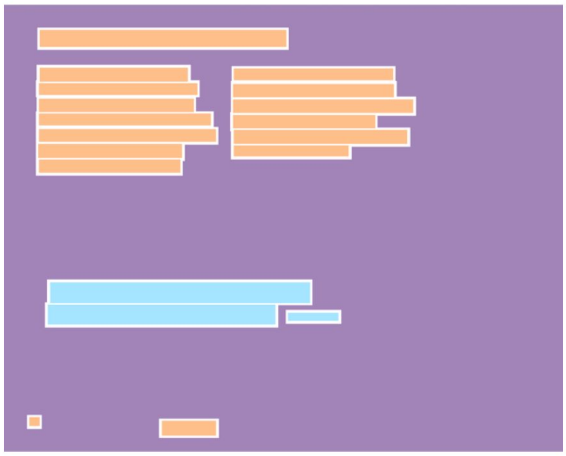


Abb. 9: Ground Truth der Seite aus Abb. 8

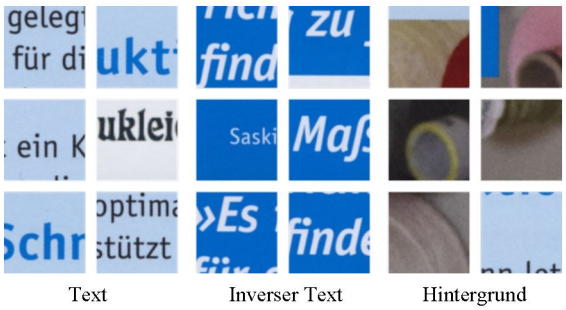


Abb. 10: Trainingsbilder

Prinzipiell ist es möglich auch ganze Seitenbilder mit Fully Convolutional Networks zu trainieren [5]. Um eine bessere Kontrolle der trainierten Seitenbereiche zu erhalten wurde jedoch der Ansatz mit Einzelbildern gewählt. Damit lässt sich sehr einfach die Anzahl der Trainingsbilder in jeder Klasse gleich groß wählen. Auch ist es möglich, einen hierarchischen Ansatz zu implementieren. Überschreitet die Texthöhe gewisse Grenzen, so wird der extrahierte Ausschnitt mit einem passenden Faktor verkleinert. So wird der Detektor auf Text einer einheitlichen Größe trainiert.

Um eine ausreichend große Menge an Trainingsdaten zu erzeugen, sind überlappende Einzelbilder erlaubt. Trainingsbilder werden nur aus den geraden Seiten der eingescannten Dokumente erzeugt. Die ungeraden Seiten werden zur Erkennung und Berechnung der Fehlerraten des Verfahrens benutzt. Aus 36 Seiten wurden insgesamt ca. 80.000 Trainingsbilder extrahiert.

6. ERKENNUNG

Wie bereits angedeutet, kommt zur Erkennung das Fully Convolutional Network zum Einsatz. Damit wäre es theoretisch möglich, die vollständigen Seitenbilder der Erkennung zuzuführen. Bei 300 dpi würde dies jedoch selbst den Speicher von high-end Grafikkarten zum Überlaufen bringen. Daher muss eine einzelne Seite in überlappenden Kacheln von ca. 320x320 Pixeln verarbeitet werden. Durch die Unterabtastung des Netzwerkes ergibt sich eine Wahrscheinlichkeit für einen jeweils 16x16 Pixel großen Block.

Zur Erkennung größerer Schrift wird auch hier ein hierarchischer Ansatz implementiert. Die Dokumentenseiten werden dem Detektor in mehreren Verkleinerungsstufen des Faktors zwei zugeführt. Für die finale Klassenzuordnung wird die Erkennung mit der höchsten Wahrscheinlichkeit herangezogen. Abbildung 11 zeigt das Resultat für die Textseite aus Abb. 8.

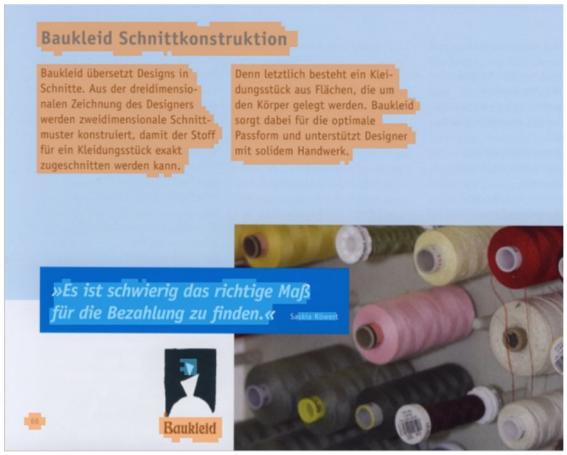


Abb. 11: Segmentierung der Seite aus Abb. 8

5. ERGEBNISSE

Auf den 36 ungeraden Seiten der verwendeten Dokumente wird die Segmentierung (Abb. 11) mit der Ground Truth (Abb. 9) verglichen. Daraus lässt sich mit verschiedenen Maßen die Güte der Erkennung quantifizieren. Tabelle 1 zeigt die Wahrheitsmatrix (Confusion Matrix) als Mittelwerte über alle Seiten. Hierbei wurde jeder erkannte Block mit der wahren Klasse des Pixels in seinem Zentrum verglichen. Tabelle 2 enthält die bei Segmentierungsfragestellungen oft verwendete mean intersection over union (mean IU). Die Berechnung erfolgt auf Pixelbasis. Dabei werden die Bereiche einer erkannten Klasse mit den Bereichen der tatsächlichen Klasse verglichen. Die mean IU ist das Verhältnis aus dem Flächeninhalt der Schnittmenge dieser Bereiche zum Flächeninhalt der Vereinigungsmenge. Die frequency weighted IU gewichtet die beteiligten Klassen nach dem Anteil ihres Auftretens. Damit wird vermieden, dass eine Klasse, die auf einer Seite nur selten auftritt (z.B. inverser Text) das Gesamtergebnis zu stark beeinflusst. Die pixel accuracy ist der Anteil aller richtig klassifizierter Pixel, während die mean accuracy diesen Anteil pro Klasse berechnet und dann über alle Klassen mittelt.

Wahr: / Erkannt:	Text	Inv. Text	Hintergr.
Text	92,8	0,1	7,1
Inverser Text	0,5	88,0	11,6
Hintergrund	0,4	0,3	99,4

***Tabelle 1:** Confusion Matrix in Prozent*

Maß	Alles	Text	Inv. Text	Hinter -grund
Pixel acc.	97,5	—	—	—
Mean acc.	96,9	98,6	94,8	97,3
Mean IU	90,7	91,6	83,9	96,7
Freq. w. IU	95,1	—	—	—

Tabelle 2: *Erkennungsgenauigkeit in Prozent*

Die Werte aus Tabelle 1 zeigen, dass der Detektor leicht dazu tendiert, Text (normal wie auch invers) als Hintergrund zu klassifizieren. Dies passiert vor allem an den Rändern von Textblöcken, wo auch ein relativ kleines Detektionsfenster noch Textanteile aufnimmt. Des Weiteren ist der hierarchische Ansatz so ausgelegt, dass er Text leicht bevorzugt. Wird in einer feineren Auflösungsstufe Hintergrund mit hoher Wahrscheinlichkeit erkannt, so kann das durch Text in einer größeren Auflösung überstimmt werden, selbst wenn die Hintergrundwahrscheinlichkeit größer ist. Sehr große Schrift sieht für die feinste Auflösungsstufe lokal oft wie homogener Hintergrund aus. Für die anschließende MRC Verarbeitung ist ein etwas zu groß erkannter Textbereich aber fast immer unproblematisch.

Die nach Klassen aufgeschlüsselten Werte aus Tabelle 2 liefern für inversen Text schlechtere Resultate als für normalen Text. Dies liegt zum einen daran, dass im betrachteten Bildmaterial weniger inverser Text zur Extraktion von Trainingsbildern vorhanden ist, zum anderen an der Tatsache, dass inverser Text oft in schwierigeren Umgebungen vorkommt. Heller Text wird oft verwendet, um Text über Bilder zu setzten.

Die Rechenzeiten hängen stark von der Anzahl und Leistungsfähigkeit der verwendeten GPUs ab. Auch deren verfügbarer Speicher fließt dabei maßgeblich ein. Er bestimmt, wie viele überlappende Kacheln pro Seite gebildet werden müssen. Auf einem Notebook mit NVIDIA GeForce GT 650M Grafikkarte mit 1 GB Speicher konnte eine DIN-A4 Seite in 300 dpi in ca. 10 Sekunden segmentiert werden. Dabei stand die Grafikkarte vollständig dem Programm zur Verfügung, um den mit 1 GB eher knapp bemessenen Speicher voll nutzen zu können.

Abschließend werden die Resultate mit einem Verfahren verglichen, das auf dem Patent [3] basiert. Bei diesem Verfahren kommen Verarbeitungsschritte zum Einsatz, die explizites Wissen um textartige Strukturen umsetzen. So wird unter anderem die Kantenaktivität berechnet, und zusammenhängende Bereiche werden nach ihrer Größe bewertet. Abbildung 13 zeigt deutlich, dass bei diesem Verfahren viele Inhalte des kontrastreichen und kantenstarken Bildes in die Maske gelangen, während bei dem hier vorgestelltem Verfahren nur sehr wenige Bereiche aus dem Bild einfließen. Insbesondere sei darauf hingewiesen, dass der über dem Bild liegende halbtransparente Text zuverlässig segmentiert wird.

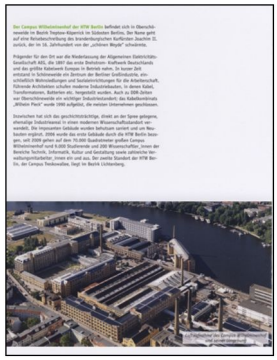
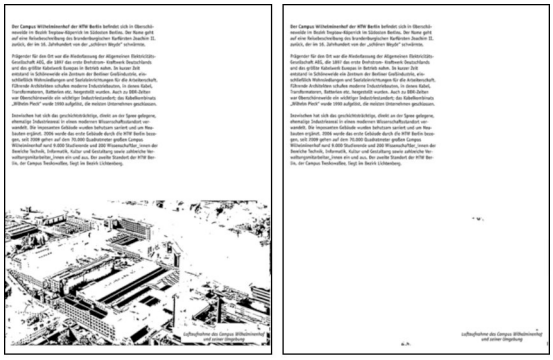


Abb. 12: *Dokumentenseite im Original*



(a) (b)

Abb. 13: *MRC Masken (a) nach Verfahren [3], (b) mit semantischer Segmentierung*

6. ZUSAMMENFASSUNG UND AUSBLICK

Es wird ein Verfahren vorgestellt, das eingescannte Dokumentenseiten auf einen 16x16 Pixel großem Raster in die Klassen Text, inversen Text und Hintergrund segmentiert. Zum Training des Detektors werden Convolutional Neural Networks eingesetzt, während die Erkennung durch Fully Convolutions Networks geschieht. Zu diesem Zwecke wird ein an GoogLeNet angelehntes, stark vereinfachtes Netzwerk entworfen und

mit ca. 80.000 Bildausschnitten aus 36 Dokumentenseiten trainiert. Die Leistungsfähigkeit der Erkennung wird auf anderen 36 Dokumentenseiten getestet, mit den gängigen Maßen bewertet und visuell mit einem Verfahren ohne neuronale Netzwerke verglichen. Die erzielten Erkennungsraten im Bereich von 80 – 95% sind als sehr gut zu bewerten, wobei es für den Dokumentenbereich nach Wissen des Autors bislang keine vergleichbaren Untersuchungen gibt. Das Verfahren zeichnet sich gegenüber anderen Ansätzen dadurch aus, dass kaum Textanteile in fotorealistischen Bildern detektiert werden.

Kritisch betrachtet wäre einzuwenden, dass das vorgestellte Verfahren nur das tut, was eine Optical Character Recognition (OCR) bereits kann. Dazu lässt sich anmerken, dass auch für ein OCR-Verfahren eine gute Segmentierung die Voraussetzung einer guten Erkennung ist. Darüber hinaus ist der vorgestellte Ansatz gegenüber einer OCR wesentlich genereller. Das zeigt sich u.a. daran, dass inverser Text ohne Mühe von normalem Text unterschieden wird, was eine OCR i.d.R. nicht leistet. Es ist auch nicht notwendig, an irgendeiner Stelle des Verfahrens eine Definition von inversem Text in den Algorithmus einzubauen. Daraus leitet sich die Erwartung ab, dass mit dem vorgestellten Ansatz auch andere Unterscheidungen im eingescannten Bildmaterial vorgenommen werden können, wie sie für eine noch genauere Klassifizierungen oder die Verschlagwortung der Inhalte notwendig wären. Dafür könnte es notwendig werden, die Komplexität des verwendeten Netzwerks wieder zu erhöhen. Der alles entscheidende Punkt ist jedoch das Bereitstellen von Trainingsdaten in sehr großer Menge, was sich in vielen Anwendungsbereichen nicht immer realisieren lässt.

5. LITERATURHINWEIS

1. Oettler, Alexandra: *PDF/A kompakt 2.0*, Association for Digital Document Standards, 2013.

2. International Telecommunication Union: Recommendation T.44: Mixed Raster Content, 2005.

3. Patent EP1104916B1: Verfahren zur Kompression von gescannten Farb- und/oder Graustufendokumenten, 1999.

4. Krizhevsky, Alex; Sutskever, Ilya; Hinton, Geoffrey E.: Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems 2012*, NIPS Proceedings 2012.

5. Shelhamer, Evan; Long, Jonathan; Darrell, Trevor: Fully Convolutional Networks for Semantic Segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Xplore 2015.

6. Simonyan, Karen; Zisserman, Andrew: Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

7. Szegedy, Christian; Liu, Wei; Jia, Yangqing; Sermanet, Pierre; Reed, Scott; Anguelov, Dragomir; Erhan, Dumitru; Vanhoucke, Vincent; Rabinovich, Andrew: Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.