technische universität
dortmund

CNN-basiertes Word Spotting in einem Kriegstagebuch aus dem 1. Weltkrieg

Eine Studienarbeit (Überarbeitete Fassung)

Matthias Kasperidus
28. Mai 2018

Supervisors:
Prof. Dr.-Ing. Gernot A. Fink
Dipl.-Inf. Leonard Rothacker

Fakultät für Informatik
Technische Universität Dortmund
http://www.cs.uni-dortmund.de

INHALTSVERZEICHNIS

1 EINLEITUNG

Im Zuge der fortschreitenden Digitalisierung wächst der Wunsch von Archiven und Historikern ihre eingescannten Dokumente mit computergestützten Verfahren schneller explorieren und analysieren zu können. Für viele solcher Dokumente existieren allerdings keine Transkripte oder sonstige maschinenlesbare Metadaten zum genauen Inhalt. Eine Volltextsuche, Schlagwortsuche oder Indexierung mit Methoden für maschinenlesbare Textdokumente sind also nicht ohne weiteres möglich. Die automatisierte Erstellung eines Transkriptes z.B. mittels Optical Character Recognition oder Techniken zur Handschrifterkennung ist für historische Dokumente im Allgemeinen schwierig. Alterungsartefakte, alte Schriftarten bzw. Schreibstile und Sprache, für die es kaum Trainingsmaterial gibt, machen den erfolgreichen Einsatz bekannter Erkennungsverfahren schwierig [GSGN17].
Besonders schwierig ist die Erkennung handschriftlicher historischer Dokumente, da bereits die Variabilität im Schriftbild eines einzelnen Schreibers sehr groß sein kann. Die Unterschiede im Schriftbild verschiedener Schreiber können soweit gehen, dass sogar Experten einige Zeit benötigen, um den Text eines unbekannten Schreibers zu entziffern. Eine Möglichkeit, solche Dokumente dennoch automatisch durchsuchen zu können, ist *Word Spotting*. Beim Word Spotting geht es darum die Vorkommen eines gegebenen Anfragewortes in Dokumentenabbildern zu finden[1]. Das Ergebnis einer Word Spotting Anfrage ist eine sortierte Liste mit Vorschlägen von Bildbereichen an denen das gesuchte Wort möglicherweise vorkommt (*Retrievalliste*). Word Spotting lässt sich also als eine spezielle Form von Bildretrieval auffassen, bei dem Wortabbilder gesucht werden. Verfahren zur Lösung von Word Spotting vermeiden eine Erkennung des Textes und reduzieren so die Schwierigkeit des Problems [GSGN17].

Seit Beginn der Entwicklung von Word Spotting Verfahren für historische Dokumente, die spätestens mit der Arbeit von Manmatha et al. [MHR96] einsetzte, wurden diverse prinzipielle Lösungsstrategien und Ansätze vorgestellt, bei denen verschiedenste Methoden der Mustererkennung eingesetzt werden [GSGN17]. Der aktuelle Trend zum Einsatz von *Convolutional Neural Networks* (CNNs) im Bereich Computer Vision spiegelt sich auch in aktuellen Arbeiten zum Word Spotting [WB16], [SF18]

1 Analog lässt sich das Word Spotting Problem auch in anderen Domänen z.B. für gesprochene Sprache formulieren [RRRG89].

wider.
Für die Entwicklung, Analyse, Evaluierung und den Vergleich von Word Spotting Verfahren werden in der Literatur verschiedene historische Datensätze und entsprechende Benchmarks verwendet. Besonders aus Gründen der Vergleichbarkeit verschiedener Verfahren und methodischer Weiterentwicklungen haben sich einige Datensätze in der Literatur etabliert (siehe z.B. [SF18] Abschnitt 4.1), wobei für den Vergleich stets darauf geachtet werden muss, dass das gleiche Evaluierungsprotokoll verwendet wird.

Für die Anwendung von Word Spotting ist interessant, ob und wie einfach sich diese Verfahren auf neue Datensätze übertragen lassen. In der vorliegenden Arbeit wird daher ein Datensatz vorgestellt, der zuvor noch nicht für die Evaluierung von Word Spotting Verfahren verwendet wurde, ein Kriegstagebuch aus dem ersten Weltkrieg. Der Datensatz wurde im Rahmen der Studienarbeit für die Evaluierung von Word Spotting Verfahren aufbereitet und für die Evaluierung eines aktuellen Word Spotting Verfahrens, das auf CNNs basiert, genutzt.

Im Folgenden wird das Kriegstagebuch vorgestellt, anschließend wird auf den weiteren Aufbau der Arbeit eingegangen.

1.1 DAS KRIEGSTAGEBUCH

Bei dem Kriegstagebuch handelt es sich um das Tagebuch des deutschen Kriegsfreiwilligen Max Götz, welches er bei seinem Einsatz im ersten Weltkrieg geführt hat. In dem Tagebuch beschreibt Max Götz Tagesabläufe, Ereignisse und Eindrücke, welche er von seinem Aufbruch am 4. Dezember 1914 bis an die Frontlinie bei der französischen Stadt Souchez erlebt hat. Die Aufzeichnungen enden am 4. Januar 1915.
Das Kriegstagebuch umfasst insgesamt 42 eingescannte Seiten eines linierten Notizbuches, wobei zwei Bilder den Einband zeigen und zwei Seiten unbeschriftet sind. Die restlichen 38 Seiten enthalten jeweils Text in From von Kurrentschrift, die zu Beginn des 20. Jahrhunderts in Deutschland verbreitet war. Die Schrift zeichnet sich durch den stark kursiven Charakter sowie lange Ober- und Unterlängen aus, die im Kriegstagebuch in die benachbarten Zeilen, teilweise sogar in den Schriftbereich anderer Worte, reichen. Die Abbildung 1.1.1 zeigt zwei Beispielseiten des Tagebuchs. Fremdeinflüsse wie Flecken oder Linien, die nicht zur Schrift gehören, stören das Schriftbild. Auffällig ist, dass die Strichstärken der Schrift, vermutlich bedingt durch das Schreibgerät, innerhalb einer Seite variieren. Außerdem zeigen die Beispielbilder, dass die Farbkontraste zwischen Schrift und Hintergrund über verschiedene Seiten variieren. Obwohl das Tagebuch von einem Schreiber stammt, unterscheidet sich das Schriftbild gleicher Worte dadurch stark.

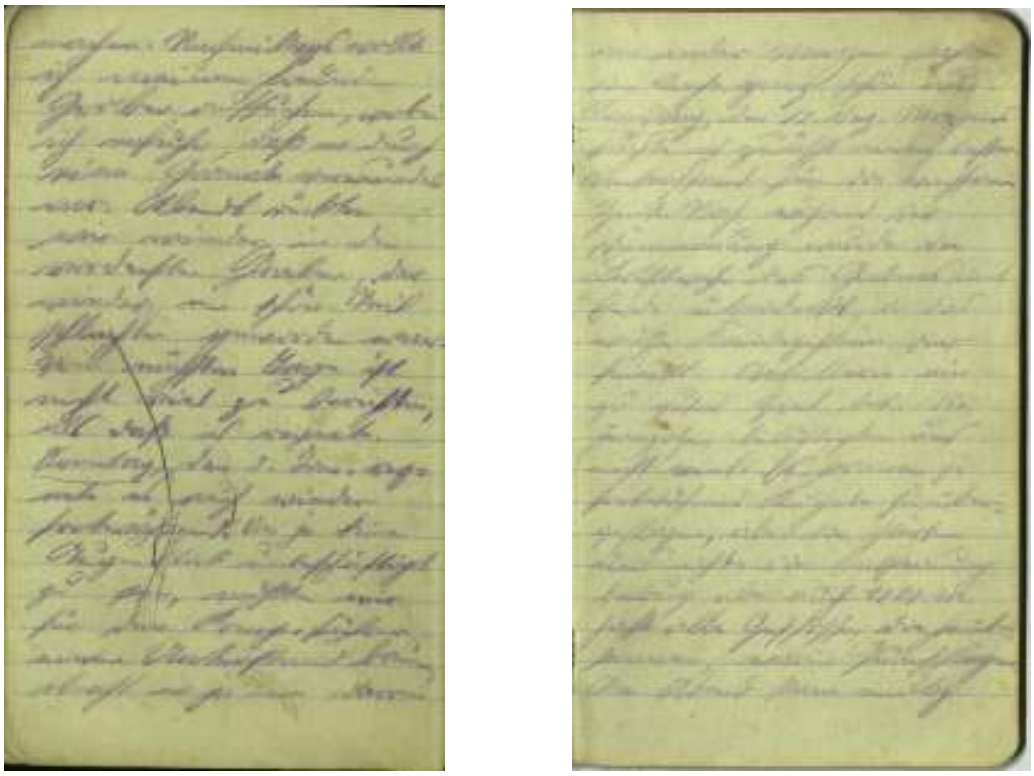


Abbildung 1.1.1: Zwei Seiten des Kriegstagebuch.

Das Kriegstagebuch stammt aus einer Privatsammlung mit Dokumenten aus dem ersten Weltkrieg, die in Dortmund (Deutschland) von der Historikerin Dr. phil. Britta Bley gepflegt wird. Die eingescannten Bilder des Kriegstagebuchs wurden zusammen mit einem Transkript von Dr. phil. Britta Bley und Prof. Dr.-Ing. Gernot A. Fink für diese Arbeit zur Verfügung gestellt.

1.2 AUFBAU DER ARBEIT

Kapitel 2 führt in das Thema Word Spotting ein. Der Fokus liegt dabei auf einer Unterteilung und Benennung prinzipieller Charakteristika von Word Spotting Verfahren sowie auf der Auswahl der CNN-basierten Methodik für das Word Spotting im Kriegstagbuch. Eine ausführliche Darstellung der benötigten Grundlagen zu CNNs findet sich in entsprechenden Lehrbüchern z.B. [GBC16]. In Kapitel 3 wird auf die Me-

thodik mit Hinblick auf die Nachvollziehbarkeit der Experimente bei der Evaluierung eingegangen. Kapitel 4 stellt den Hauptteil dieser Studienarbeit dar. In dem Kapitel wird zuerst auf die Vorbereitung des Kriegstagebuchs für die Evaluierung zum Word Spotting eingegangen. Dazu gehört die Erstellung von *Bounding Box* Annotation für die Wörter des Kriegstagebuchs sowie die Festlegung eines Benchmarks. Im zweiten Teil von Kapitel 4 wird auf die Experimente eingegangen, die zur Beurteilung der Word Spotting Methodik im Bezug auf das Kriegstagebuch dienen. Abschließend werden die wichtigsten Resultate in Kapitel 5 zusammengefasst.

2 WORD SPOTTING

Grundsätzlich läuft das Retrieval beim Word Spotting immer über eine Ähnlichkeitsbewertung von Bildregionen bezüglich der Anfrage ab [GSGN17]. Dazu werden zunächst Merkmalsrepräsentationen für Bildregionen berechnet, die Ähnlichkeitsbewertung findet dann auf Basis dieser Merkmalsrepräsentationen statt. Anschließend lassen sich bewertete Bildregionen nach Ähnlichkeit sortieren und als Retrievalliste zurückgegeben [GSGN17].

Im ersten Teil dieses Kapitels wird auf Unterscheidungsmerkmale von Word Spotting Methoden eingegangen. Dabei wird auch die Auswahl der CNN-basierten Methodik zum Word Spotting im Kriegstagebuch begründet. Im zweiten Teil dieses Kapitels wird die Merkmalsrepräsentation der ausgewählten Methodik besprochen die maßgeblich durch andere Arbeiten zum Word Spotting geprägt wurde. Außerdem erfolgt eine Benennung weiterer Arbeiten zu CNN-basiertem Word Spotting.

TEIL 1 Im Bereich der Analyse von historischen Dokumentenabbildern wurde der Begriff Word Spotting durch Arbeiten von Manmatha et al. [MHR96] populär. In ihrer Arbeit ist die Word Spotting Idee direkt mit einer Anwendung von Word Spotting, der Indexierung von Dokumenten, verbunden. Dazu werden zunächst Wortabbilder segmentiert und anschließend automatisch nach visueller Ähnlichkeit gruppiert. Den Gruppen werden anschließend manuell entsprechende Schlüsselworte zur Indexierung zugeordnet. Der Index kann dann für das Durchsuchen der Dokumente nach diesen Schlüsselworten genutzt werden. Das in der Einleitung genannte gängige Verständnis von Word Spotting als *Retrieval* von Wortabbildern ohne eine explizite Erkennung schließt den Anwendungsfall der Indexierung mit ein (vgl. [GSGN17] Abschnitt 1.1). In [GSGN17] werden Word Spotting Methoden in drei grundlegenden Punkten unterschieden.

Der erste Punkt ist die Art der verwendeten Anfrage. Am bekanntesten und relevant für die vorliegende Arbeit sind die Anfragearten *Query-by-Example* und *Query-by-String*. Im Query-by-Example Fall (QbE) dient ein Beispielbild des Anfragewortes als Eingabe. Für den Anwender ergibt sich der Nachteil, dass dieser erst ein Beispielbild des Anfragewortes in den Dokumenten finden muss. Ein Vorteil vieler QbE Verfahren ist, dass sie ohne annotiertes Trainingsmaterial auskommen [GSGN17]. Im Query-by-

String Fall (QbS) wird das Anfragewort als maschinenlesbare Zeichenkette (*String*), z.B. im ASCII Format, angegeben. Dokumentenkollektionen können so nach beliebigen Wörtern durchsucht werden, ohne dass ein Beispielbild angegeben werden muss. Ein Nachteil vieler QbS Verfahren ist allerdings, dass sie annotiertes Trainingsmaterial benötigen [GSGN17].

Der zweite Punkt ist die Unterscheidung zwischen Word Spotting Methoden, welche ein vorheriges Training mit annotiertem Trainingsmaterial benötigen (*trainingsbasiert*) und solchen die keines benötigen (*trainingsfrei*). Sofern genügend annotiertes Trainingsmaterial verfügbar ist liefern trainingsbasierte Verfahren bessere Word Spotting Ergebnisse als trainingsfreie Verfahren [GSGN17]. Im Fall von historischen Dokumenten kann die Menge an annotiertem Trainingsmaterial, wie in der Einleitung erwähnt, gering und damit problematisch für trainingsbasierte Verfahren sein. Für einen CNN-basierten Ansatz, nämlich den des PHOCNet [SF16], wurde jedoch gezeigt, dass ein vorheriges Training (*Pretraining*) auf synthetisch erzeugten Daten in Kombination mit einer Adaptierung durch eine geringe Menge an spezifischem Trainingsmaterial Ergebnisse auf historischen Datensätzen erreichen kann, die dem Stand der Technik (*State-of-the-Art*) entsprechen [GSF18]. Diese Erkenntnis und die Tatsache, dass der PHOCNet-basierte Ansatz den State-of-the-Art zum *segmentierungsbasierten* Word Spotting auf einer Reihe von (historischen) Datensätzen darstellt [SF18], begründen die Verwendung des PHOCNet in der Methodik der vorliegenden Arbeit (siehe Kapitel 3.1). Dies führt zum dritten Unterscheidungspunkt von Word Spotting Methoden.

Die dritte Unterscheidung besteht zwischen *segmentierungsbasierten* und *segmentierungsfreien* Verfahren. Segmentierungsbasierte Verfahren gehen davon aus, dass die Dokumentenabbilder in einem Vorverarbeitungsschritt in einzelne Wort- oder Zeilenabbilder segmentiert wurden [GSGN17]. Für das Retrieval werden anschließend nur noch diese Bildausschnitte berücksichtigt. Im Falle des PHOCNet-basierten Ansatzes wird von einer Wortsegmentierung ausgegangen [SF16], daher wird auch im restlichen Teil der vorliegenden Arbeit davon ausgegangen, wenn von segmentierungsbasiertem Word Spotting die Rede ist. Für die Anwendung des PHOCNet-basierten Ansatzes im Kriegstagebuch, wie er in [SF16] vorgestellt wurde, wäre also auch der Vorschlag und die Auswertung einer Segmentierungsmethode für Wortabbilder nötig. Gerade bei historischen Dokumenten ist eine Wortsegmentierung aus in der Einleitung genannten Gründen jedoch häufig schwierig [GSGN17], was die segmentierungsfreien Word Spotting Methoden begründet. Segmentierungsfreie Verfahren verzichten auf eine solche Segmentierung und ziehen bei der Anfrage zunächst deutlich mehr Bildregionen für die Ähnlichkeitsbewertung in Betracht. In der Literatur finden sich segmentierungsfreie Ansätze die zwischen zwei verschiedenen Herangehensweisen variieren.

Die eine Herangehensweise ist eine Auswahl von Bildregionen für das Retrieval, die ausschließlich auf Basis von Informationen über die gestellte Anfrage getroffen wird. Beispiele sind Patch-basierte Ansätze [RATL11], [RRF13], die Bildregionen in der Größe des Anfragewortes in einem dichten Gitter über die gesamten Dokumentenseiten betrachten. Die Merkmalsrepräsentationen für die Ähnlichkeitsbewertung ergeben sich dabei erst zur Anfragezeit, was zu einem hohen Berechnungsaufwand zur Anfragezeit führt. Dafür werden kaum Annahmen über die Schrift und Struktur der Dokumentenabbilder gemacht, was einen vorzeitigen Ausschluss relevanter Bildregionen vermeidet.
Die andere Herangehensweise ist eine Vorauswahl von Bildregionen für das Retrieval, die ausschließlich auf Grundlage der zu durchsuchenden Dokumente getroffen wird. Beim Retrieval wird dann nur noch die Vorauswahl der Bildregionen betrachtet. Von segmentierungsfreiem Word Spotting wird dann gesprochen, wenn es sich bei der Vorauswahl um eine Segmentierung handelt, bei der bewusst Segmentierungsfehler zugelassen werden [RSR+17]. Es wird also Angenommen, dass es zu Über- und Untersegmentierungen kommen kann. In [RSR+17] nennt die Bildregionen einer solchen fehlerhaften Segmentierung *Worthypothesen*. Ein Nachteil dieser Herangehensweise ist der vorzeitige Ausschluss relevanter Bildregionen durch die möglichen Untersegmentierungen, diese sollten daher möglichst vermieden werden. Ein Vorteil solcher Verfahren ist, dass sich Merkmalsrepräsentationen der Worthypothesen Vorberechnen lassen, was das Word Spotting zur Anfragezeit effizient machen kann [RSR+17].
Technisch gesehen lässt sich jedes segmentierungsbasierte Verfahren durch Worthypothesen zu einem segmentierungsfreien Verfahren umrüsten. Es bleibt allerdings die Frage, wie gut das segmentierungsbasierte Verfahren bei der Ähnlichkeitsbewertung mit Segmentierungsfehlern umgehen kann. Für den PHOCNet-basierten Ansatz zusammen mit Worthypothesen wurden in [RSR+17] bei drei historischen Datensätzen segmentierungsfreie Word Spotting Ergebnisse erreicht, die dem State-of-the-Art entsprechen oder vergleichbar mit diesem sind. Für das segmentierungsfreie Word Spotting im Kriegstagebuch wird daher die Methodik aus [RSR+17] verwendet (siehe Kapitel 3).

TEIL 2 Der Erfolg der ausgewählten Methodik lässt sich nicht alleine durch die Verwendung von Convolutional Neural Networks begründen. Entscheidend ist auch die Idee der Merkmalsrepräsentation, welche dem PHOCNet zugrunde liegt. Der Name PHOCNet leitet sich aus der von Almazan et al. in [AGFV14] vorgeschlagenen Attributrepräsentation dem *Pyramidal Histogram of Characters* (PHOC) ab. Die PHOC Repräsentation ist natürlicherweise für Strings definiert und ist ein Binärvektor, der auf mehreren Hierarchieebenen angibt, welche Buchstaben eines vorgegebenen Alphabets

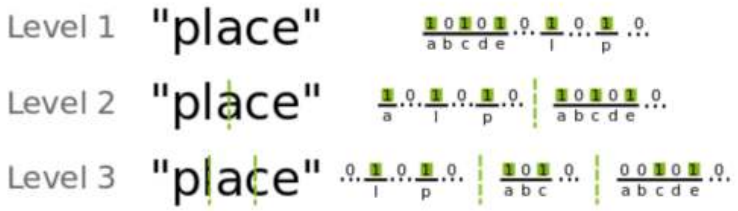


Abbildung 2.0.1: Beispiel der Bildung eines PHOC-Vektors bei drei Hierarchieebenen (Level), mit jeweils 1, 2 und 3 Abschnitten. Die Konkatenation der Binärvektoren jeder Hierarchieebenen bildet den PHOC-Vektor. Abbildung entnommen aus [SF16].

in bestimmten Abschnitten eines Strings vorkommen. Die Anzahl der Hierarchieebenen und die Anzahl der Abschnitte jeder Ebene können prinzipiell variiert werden. Die Abbildung 2.0.1 zeigt an einem Beispiel wie der PHOC-Vektor eines Wortes gebildet wird. Nach einer Transformation von Wortabbildern und Anfrage in einen gemeinsamen PHOC-Vektorraum (*PHOC-Embedding*) lässt sich die Ähnlichkeitsbewertung für das Retrieval durch Distanzberechnungen in diesem Vektorraum durchführen. So sind insbesondere Query-by-Example und Query-by-String Anfragen möglich. In [AGFV14] werden mehrere SVMs für die Transformation von Wortabbildern in einen PHOC-Vektorraum verwendet. Das PHOCNet als CNN-basierte Transformation übertrifft die SVM-basierte Transformation bezüglich der Word Spotting Qualität [SF16]. Neben der PHOC Repräsentation wurden auch andere Attributrepräsentationen in der Literatur zum Word Spotting vorgeschlagen, die sich mit der grundlegenden Architektur des PHOCNet berechnen lassen [SF17]. Ein Vergleich dieser Attributrepräsentationen in [SF17] im segmentierungsbasierten Word Spotting Szenario auf mehreren Datensätzen ergab jedoch, dass keine dieser Repräsentationen eine allgemeine Überlegenheit gegenüber den Anderen aufweist. Die Attributrepräsentation, die in der vorliegenden Arbeit verwendet wird, ist daher die PHOC Repräsentation.

Eine Übersicht über weitere Arbeiten die CNN-basiertes Word Spotting thematisieren findet sich in [SF18]. Zwei Veröffentlichungen zu segmentierungsfreiem CNN-basiertem Word Spotting, die nicht in [SF18] referenziert werden, sind [GV17] und [WLB17]. Beide Ansätze nutzen die CNNs ebenfalls zur Vorhersage von Attributrepräsentationen, insbesondere zu der von PHOC Repräsentationen. Außerdem basieren beide Ansätze prinzipiell auf einer Vorauswahl von Bildregionen als Worthypothesen und eine entsprechende Vorberechnung der Attributrepräsentationen, ähnlich wie [RSR+17] .

3 METHODIK

In diesem Kapitel wird die in Kapitel 2 ausgewählte CNN-basierte Methodik zum Word Spotting im Krigstagebuch (siehe Kapitel 1.1) beschrieben. Die Methodik stammt aus [RSR+17] und besteht im Wesentlichen aus zwei Komponenten:

Die erste Komponente erzeugt eine Menge von Worthypothesen auf der Basis sogenannter *Textdetektoren*, die Bewertungen zur „Texthaftigkeit" einzelner Pixel bzw. kleinerer Bildregionen vornehmen. In Anlehnung an [RSR+17] werden diese Bewertungen im Folgenden als *Textscores* bezeichnet. Die Worthypothesen werden einmal initial unabhängig vom Anfragewort je Dokumentenseite erzeugt, wobei Textscores von verschiedenen Textdetektoren verwendet werden. Abschnitt 3.2 dieses Kapitels geht genauer auf die Erzeugung der Worthypothesen ein.

Die zweite Komponente berechnet Merkmalsrepräsentationen der Worthypothesen und der Anfrageworte. Wie in [RSR+17] wird in der vorliegenden Arbeit der PHOCNet-basierte Ansatz von Sudholt et al. [SF17] zur Berechnung von Attributrepräsentationen verwendet. Das als TPP-PHOCNet bezeichnete tiefe CNN wird im segmentierungsbasierten Szenario für die Berechnung von PHOC-Vektoren trainiert. In der vorliegenden Arbeit werden wie in [SF17] PHOC-Vektoren mit 5 Stufen und einer Unterteilung in jeweils $1, 2, 3, 4$ bzw. 5 Abschnitte verwendet. Die Berechnung der PHOC-Vektoren muss für die Worthypothesen ebenfalls nur einmal initial nach dessen Erzeugung durchgeführt werden. Abschnitt 3.1 geht genauer auf die Architektur und das Training des TPP-PHOCNet ein.

Zur Anfragezeit wird zuerst der PHOC-Vektor des Anfragewortes berechnet. Im Query-by-String Szenario ergibt sich der PHOC-Vektor des Anfragewortes direkt, im Query-by-Example Szenario wird das trainierte TPP-PHOCNet für die Berechnung verwendet. Das Retrieval wird anschließend über die Berechnung der Kosiunsdistanzen zwischen den PHOC-Vektoren von Worthypothesen und Anfrage PHOC-Vektor durchgeführt. Damit nicht bei jeder Anfrage alle Worthypothesen verglichen werden müssen wird die gleiche Pruning-Heuristik wie in [RSR+17] verwendet. Alle Worthypothesen deren Seitenverhältnis um mehr als das 5-Fache vom Seitenverhältnis der Anfrage abweicht werden verworfen. Für Query-by-String Anfragen wird das Seitenverhältnis aus einer Schätzung der durchschnittlichen Buchstabenbreiten und Worthöhen ermittelt. Abschließend wird eine *Non-maximum supression* je Dokumentenseite durchgeführt, bei der nur die Worthypothesen in der Retreivalliste behalten

werden, die sich nicht zu mehr als 1% Intersection over Union (IOU) (siehe [NB17] Kapitel 4.1.1) mit einer ähnlicher bewerteten Worthypothese überlappen. Die nach Ähnlichkeiten sortierte Liste der verbleibenden Worthypothesen wird als Ergebnis für die Anfrage zurückgegeben.

Im Folgenden wird auf die beiden Kernkomponenten der Methodik eingegangen.

3.1 TPP-PHOCNET

Das TPP-PHOCNet [SF17] ist eine Weiterentwicklung des PHOCNet [SF16] und wurde speziell zur Berechnung von Attributrepräsentationen für Wortabbilder entwickelt. Der wesentliche Unterschied des PHOCNet im Vergleich zu klassischen Convolutional Neural Networks zur Klassifikation ist der *Sigmoid-Layer* zur Ausgabe eines Attributvektors am Ende des PHOCNets [SF16]. Jedes Ausgabeneuron ist unabhängig von den anderen Ausgabeneuronen für die Vorhersage eines Attributes zuständig. Im Falle der PHOC Repräsentationen soll der PHOC-Vektor des Wortes eines eingegebenen Wortabbildes vorhergesagt werden. Es sei angemerkt, dass es sich bei der Ausgabe des PHOCNet wegen der Verwendung der Sigmoid Funktion nicht um einen Binärvektor handelt. Die Einträge der vorhergesagten PHOC-Vektoren nehmen stattdessen Werte im Intervall $[0, 1]$ an.
Für den Umgang mit variablen Eingabebildgrößen ist ein *Spatial Pyramid Pooling Layer* (SPP-Layer) [HZRS14] im ursprünglichen PHOCNet verwendet worden. Im TPP-PHOCNet wurde der SPP-Layer durch den *Temporal Pyramid Pooling Layer* (TPP-Layer) [SF17] ersetzt. Der TPP-Layer ist eine spezielle Anpassung des SPP-Layers, bei dem lediglich eine horizontale Unterteilung erfolgt. Damit soll der sequentielle Charakter der Schrift besser berücksichtigt werden [SF17]. Für die Experimente der vorliegenden Arbeit wurde, wie in [SF17], ein TPP Layer mit 5 Hierarchieebenen und jeweils $1, 2, 3, 4$ bzw. 5 Abschnitten verwendet.

Das Training der Netzwerkparameter des TPP-PHOCNet beruht, wie für CNNs üblich, auf Gradientenverfahren zur Optimierung einer *Loss-Funktion*. Als Loss-Funktion wird wegen der binären PHOC Repräsentation, die für das PHOCNet Training übliche Loss-Funktion der *Binary Cross Entropy Loss*, auch *Sigmoid Cross Entropy Loss* genannt, verwendet [SF18]. Optimiert wird die Loss-Funktion wie in [SF16] mittels *Stochastic Gradient Descent* (SGD). Sofern nicht anders bei den Experimenten der vorliegenden Arbeit angegeben werden die gleichen SGD Parameterwerte wie in [SF16] III.C verwendet. Zu erwähnen sind die *Lernrate* und die *Anzahl der Trainingsiterationen*, die bei manchen Experimenten der vorliegenden Arbeit verändert wurden. Die Standardeinstellungen dieser Parameter sind 80000 Trainingsiteration, wobei die ersten

70000 Iterationen mit einer Lernrate von 10^{-4} und die restlichen 10000 Iterationen mit einer Lernrate von 10^{-5} durchgeführt werden. Die Methode zur Initialisierung der Netzwerkparameters folgt ebenfalls [SF16], sofern kein vortrainiertes TPP-PHOCNet Modell als Ausgangspunkt für das Training genutzt wird.

Ein essentieller Teil für den Erfolg tiefer CNNs im Allgemeinen ([GBC16] Kapitel 7) und auch für den des PHOCNets ist die Verwendung von Regularisierungstechniken [SF16]. Zum Training des TPP-PHOCNet in der vorliegenden Arbeit werden in allen Fällen *Dropout*, also das zufällige Ausblenden eines gewissen Anteils der Neuronen im CNN, und eine *Datenaugmentierung*, also die künstliche Erweiterung der Trainingsmenge durch Transformationen der Bilder aus der ursprünglichen Trainingsmenge, wie in [SF16] verwendet.

Die verwendete Implementierung des TPP-PHOCNet basiert auf dem CNN Framework Caffe [JSD+14] und wurde auch in [SF17] verwendet.

3.2 WORTHYPOTHESEN

Die Erzeugung der Worthypothesen basiert auf der Idee der *Extremal Regions* aus [MCUP04] in Kombination mit den Textscores von Textdetektoren [RSR+17]. Extremal Regions (ERs) sind zusammenhängende Bildbereiche, in denen alle Pixel größere Werte als die benachbarten Pixel besitzen [MCUP04]. Bei diesen Werten kann es sich um Grauwerte wie in [MCUP04] oder auch um Textscores wie in [RSR+17] handeln. Die Annahme bei der Erzeugung der Worthypothesen ist, dass Wörter durch solche ERs repräsentiert sind, d.h. dass sich der Bildbereich der Wörter durch niedrigere Textscores vom Hintergrund und anderen Wörtern abgrenzt. *Bounding Boxen*, welche einzelne ERs einschließen, werden dann als Worthypothesen verwendet.
Für einen endlichen Wertebereich bzw. eine Liste von Schwellwerten lässt sich die Menge aller ERs eines Bildes effizient [MCUP04] als Baum von Zusammenhangskomponenten [RSR+17] berechnen. Die Ebenen des Baumes entsprechen von der Wurzel aus aufsteigenden den Schwellwerten und enthalten jeweils die Zusammenhangskomponenten (ERs) innerhalb des Dokumentenabbildes, welches mit dem entsprechenden Schwellwert binarisiert wurde.
Anstatt alle ERs für die Worthypothesen zu verwenden, werden nur solche ERs betrachtet, welche Geschwister im Baum besitzen [RSR+17], da diese durch lokale Minima in den Textscores getrennt werden. Es wird angenommen, dass solche lokalen Minima häufig Wortzwischenräume und Zeilenabstände repräsentieren [RSR+17]. Durch die einstellbare Anzahl an Schwellwerten (*ER-Schwellwerte*), die den Wertebereich der Textscores in äquidistante Bereiche unterteilen, kann reguliert werden wie „tief“ diese

Minima maximal sein dürfen. Je mehr ER-Schwellwerte verwendet werden, desto kleinere und desto mehr lokale Minima werden tendenziell detektiert, wodurch die Anzahl der Worthypothesen steigt. Die beschriebene Auswahlheuristik von ERs (*ER-Heuristik*) wurde speziell für die Generierung von Worthypothesen entworfen [RSR+17] und unterscheidet sich damit von der Auswahlheuristik des MSER-Detektors [MCUP04].

In [RSR+17] wurde festgestellt, dass eine Höhenquantisierung der Worthypothesen zu besseren Word Spotting Ergebnissen führen kann, daher wird auch in der vorliegenden Arbeit eine Höhenquantisierung der Worthypothesen verwendet. Eingestellt wurde die Höhenquantisierung auf ein Intervall von 45 bis 400 Pixel mit einer Schrittweite von 5 Pixeln, wie in [RSR+17]. Die Einstellung des Intervalls erfolgte nach einer ersten Ansicht der Validierungsseiten des Benchmarks zum Kriegstagebuch (siehe Kapitel 4.2), bei der festgestellt wurde, dass die Worthöhen bedingt durch die Zeilenlinien des Notizbuches etwa in diesem Intervall liegen.

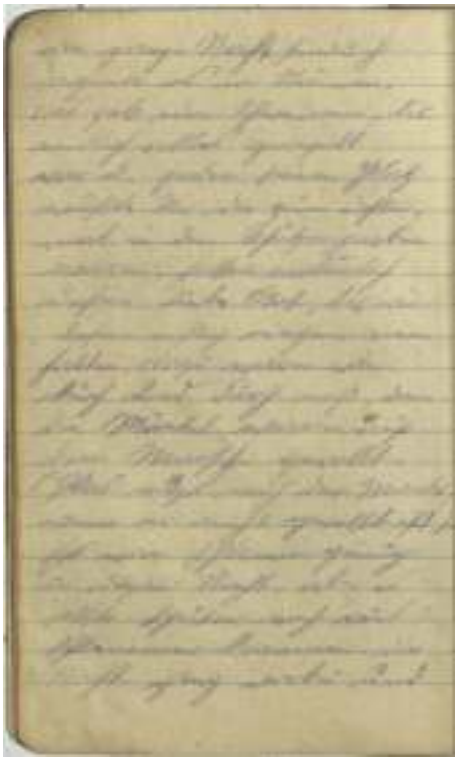
In der vorliegende Arbeit wurden zwei der in [RSR+17] vorgeschlagenen Textdetektoren für die Generierung von Worthypothesen genutzt.
Bei dem ersten Textdetektor handelt es sich um den *SIFT-Kontrast Textdetektor*, welcher Kontrastwerte von SIFT-Deskriptoren [Low04], die in einem *dichten Gitter* berechnet werden, als Textscores verwendet [RSR+17]. Der Textdetektor basiert also auf der naheliegenden Annahme, dass Schrift durch kontrastreiche Bildregionen repräsentiert ist. Eine Einstellungen der wichtiger Parameterwerte des SIFT-Kontrast Textdetektors, das sind die *Zelleneinteilung* und *Zellengröße* des SIFT-Deskriptors sowie die *Deskriptorabstände* im dichten Gitter [RSR+17], erfolgt bei den Experimenten in Kapitel 4.4.2.
Bei dem anderen Textdetektor handelt es sich um den Attribute Activation Map Textdetektor (*AAM Textdetektor*), der auf einer Modifikation des PHOCNet, dem AAM-PHOCNet [RSR+17], basiert. Das AAM-PHOCNet wurde von den Class Activation Maps in [ZKL+16] inspiriert. Die Idee des AAM-PHOCNet ist es vorherzusagen, welchen Einfluss einzelne Bildregionen auf die Vorhersage der verschiedenen PHOC-Attribute haben. Dazu wird das PHOCNet in [RSR+17], ähnlich wie in [ZKL+16], zu einem Fully Convolutional Neural Network [LSD15] modifiziert. Die letzte Netzwerkschicht des AAM-PHOCNet ist dann ein *Convolutional-Layer*, der einen Filter je PHOC-Attribut besitzt. Die Ausgabe des AAM-PHOCNet ist dann eine *Feature-Map*, welche pixelweise die Aktivierungswerte für entsprechende Vorhersagen der PHOC-Attribute enthält. Da die PHOC-Attribute das Vorkommen einzelner Textbausteine beschreiben nutzt der AAM Textdetektor die jeweils größte Aktivierung pro Pixel der Ausgabe des AAM-PHOCNet als Textscore. Das Training des AAM-PHOCNet erfolgt wie das des TPP-PHOCNet mit segmentierten Wortabbildern, wobei am Ende des AAM-PHOCNet ein *Global Average Pooling Layer* [ZKL+16] und ein Sigmoid-Layer ergänzt werden (siehe auch [RSR+17]).

4 EVALUIERUNG

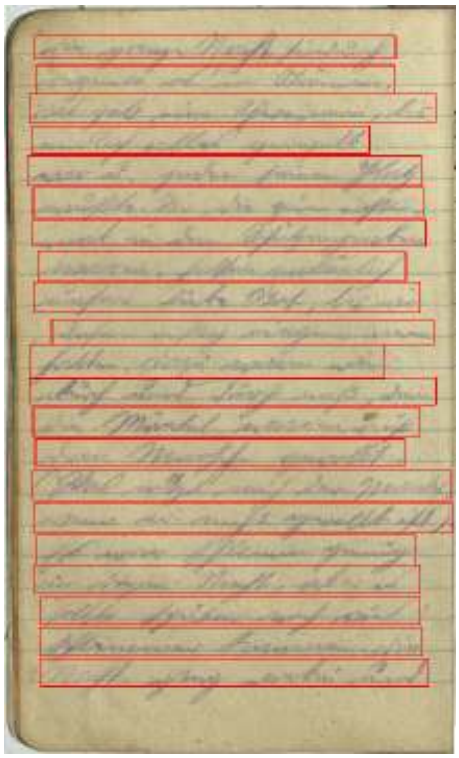
Zur Evaluierung wurden Bounding Box Annotationen für fast alle Wörter des Kriegstagebuchs erstellt. Zudem ist das Tagebuch seitenweise in einen Trainings-, Validierungs- und Testteildatensatz partitioniert worden. Zusammen mit Listen von Anfragewörten bilden die Partitionierung und Wortannotationen die Datenbasis für einen Benchmark zu unterschiedlichen Word Spotting Szenarien: Segmentierungsbasiert und segmentierungsfrei, jeweils mit Query-by-String (QbS) und Query-by-Example (QbE) Anfragen.
Das Evaluierungskapitel ist wie folgt aufgebaut:
In Abschnitt 4.1 wird auf die Erstellung der Wortannotationen des Kriegstagebuchs eingegangen. Danach erfolgt eine Beschreibung des Benchmarks (4.2), der in den anschließenden Experimenten zur Auswertung der Word Spotting Methodik verwendet wird. Inspiriert von der Arbeit von Gurjar et al. [GSF18] wird in einigen Experimenten zusätzliches Trainingsmaterial verwendet. Abschnitt 4.3 geht auf dieses zusätzliche Trainingsmaterial ein. Im letzten Abschnitt werden die Experimente besprochen, die darauf ausgerichtet sind ein, bezogen auf den Benchmark (siehe Abschnitt 4.2), möglichst gutes segmentierungsfreies Word Spotting mit der vorgestellten Methodik im Kriegstagebuch zu erreichen.

4.1 ANNOTATION

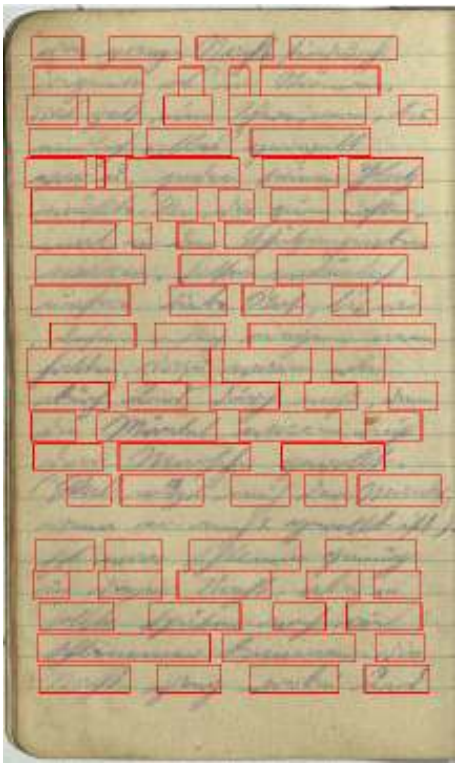
Die Erstellung der Wortannotationen erfolgte auf Grundlage des Transkriptes zum Kriegstagebuchs (siehe Kapitel 1.1). Zur Beschleunigung des Annotationsverfahrens wurde ein Forced-Alignment wie in [SRF17] durchgeführt. Das Forced-Alignment nutzt die *Bag-of-Features HMM*s aus [RF16], welche sich durch die Verwendung der *Bag-of-Features* basierten Merkmalsrepräsentation und den Verzicht auf heuristische Vorverarbeitungsschritte, wie z.B. eine Binarisierung, besonders für den Einsatz in historischen Dokumenten eignen.
Die benötigten Zeilenannotationen wurden vorher durch manuelle Markierung der Zeilen, durch Bounding Boxen, erstellt. Bei der Markierung der Zeilen wurden Ober- und Unterlängen der Schrift teilweise abgeschnitten, damit das Bild einer Zeile nicht zu sehr durch die Schrift anderer Zeilen gestört wird. Das Transkript ist bereits in einzelne Teiltranskripte je Seite unterteilt und jedes Seitentranskript enthält Zeile-



(a) Seite 16 des Kriegstagebuchs



(b) Manuell erstellte Bounding Boxen für die Zeilen



(c) Wortannotationen nach dem Forced-Alignment

(d) Manuell nachgebesserte Wortannotationen

Abbildung 4.1.1: Dargestellt ist die Seite 16 des Kriegstagebuchs mit den Zwischenergebnissen der Schritte des im Text beschriebenen Annotationsprozesses.

	#Text-/Seiten	#Textzeilen	#Wörter
Train	12/13	252	950
Val	6/8	125	484
TrainVal	18/21	377	1434
Test	20/21	390	1453

Tabelle 4.2.1: Details zu den Teildatensätzen des Benchmarks. Angegeben sind die Anzahlen der Seiten, der annotierten Textzeilen und der annotierten Wörter (siehe 4.1).

numbrüche entsprechend den Textzeilen des Kriegstagebuchs. Eine Zuordnung der Zeilentranskripte zu entsprechenden Bounding Boxen der Zeilen konnte daher automatisch durchgeführt werden. Insgesamt wurden so 767 Textzeilen des Tagebuchs annotiert. Das Kriegstagebuch enthält einige durchgestrichene Wörter, diese wurden im Zeilentranskript mit dem Unicode Ersatzzeichen (U+FFFD) markiert, um beim Forced-Alignment mit modelliert werden zu können. Das Forced-Alignment wurde anschließend genutzt, um die horizontalen Wortgrenzen innerhalb der Zeilen zu markieren. Sonderzeichen wie z.B. Satzzeichen wurden beim Forced-Alignment mit modelliert, bei der Markierung der horizontalen Wortgrenzen aber wie Leerzeichen behandelt. Mit zwei Ausnahmen (siehe Anhang A.1) sind damit keine Sonderzeichen in den Wortannotationen enthalten. Die Bounding Boxen der Wortannotationen wurden nach dem Forced-Alignment manuell nachgebessert, sodass die Wörter jeweils vollständig und möglichst exakt durch die Bounding Box erfasst sind. Abbildung 4.1.1 zeigt Zwischenergebnisse der Schritte des Annotationsprozesses beispielhaft für eine Seite des Kriegstagebuchs. Insgesamt wurden so 2887 Wörter annotiert.

4.2 BENCHMARK

Für die Evaluierung sind 21 zufällig ausgewählte Seiten des Kriegstagebuchs für Testzwecke gedacht, d.h. mit Hilfe dieser Seiten dürfen keine Parameter der Methodik eingestellt werden. Die übrigen 21 Seiten werden für das Training und die Validierung verwendet. 13 dieser Seiten sind für das Training während der Validierungsphase eingeteilt. Die anderen 8 Seiten werden zum Testen und zur Einstellung von Paramtern während der Validierungsphase genutzt. In Tabelle 4.2.1 sind Details zu diesen Teildatensätzen und deren Annotationen eingetragen. Im Folgenden werden die Teildatensätze entsprechend der Tabelle mit *Train*, *Val*, *TrainVal* und *Test* abgekürzt.

	#QbE (Seg-Free)	#QbE (Seg-Based)	#QbS (Both)
Validierung	484	267	300
Test	1453	968	679

Tabelle 4.2.2: Details zu den Anzahlen der Anfragewörter in den jeweiligen Word Spotting Szenarien.

Die Wortannotationen lassen sich für die Auswertung segmentierungsfreier und segmentierungsbasierter Word Spotting Szenarien verwenden, dabei können jeweils die beiden Anfragearten Query-by-Example (QbE) und Query-by-String (QbS) betrachtet werden. Im segmentierungsbasierten Szenario werden die Wortannotationen des jeweiligen Teildatensatzes (Val bzw. Test), auf dem ausgewertet wird, als Wortsegmentierung verwendet. Als Anfragewörter werden jeweils alle Wörter des Val bzw. Test Teildatensatzes verwendet. Im beiden QbS Szenarien werden gleiche Wörter nur einmal bei der Auswertung angefragt. Im segmentierungsbasierten QbE Szenario werden nur die Anfragewörter berücksichtigt, welche mindestens zweimal innerhalb der jeweiligen Wortannotationen vorkommen. Es ist also mindestens ein anderes Exemplar des Anfragewortes vorhanden, das in der Retrievalliste möglichst weit oben stehen sollte. Im segmentierungsfreien QbE Szenario werden auch die Wörter als Anfragen verwendet, welche nur einmal in den jeweiligen Wortannotationen vorkommen. Schließlich ist nicht klar, ob die Annotation der Anfragewörter auch durch Worthypothesen abgedeckt sind. Die genauen Anzahlen der Anfragewörter in den unterschiedlichen Word Spotting Szenarien sind in Tabelle 4.2.2 eingetragen. Bei der Auswertung wird kein Unterschied zwischen Groß- und Kleinschreibung gemacht, die Anfrageworte für das Query-by-String Szenario sind dementsprechend nur in Kleinbuchstaben vorgegeben.

Als Evaluierungsmaße werden die *Detektionsrate* (DR), der *mittlere Recall* (mR) und die *mittlere Average Precision* (mAP) verwendet. Letztere sind typische Evaluierungsmaße für die Auswertung von Retrieval Ergebnissen (siehe [MRS08] Kapitel 8) und werden daher häufig zur Auswertung in Arbeiten zum Word Spotting genutzt [GSGN17]. Die Verwendung der DR ist durch die ausgewählte segmentierungsfreie Methodik begründet. Die DR gibt das Verhältnis zwischen der Anzahl an Wortannotationen, die von mindestens einer Worthypothese zu 50% IOU überlappt werden, und der Anzahl aller Wortannotationen des entsprechenden Teildatensatzes (Val bzw. Test) an.
Der mR ist der Durchschnitt der *Recall*-Werte aller Anfragen. Der *Recall* einer Anfrage ist das Verhältnis zwischen der Anzahl relevanter Worthypothese in der Retrievalliste und allen Exemplaren des Anfragewortes in den Wortannotationen des entsprechen-

den Teildatensatzes. Eine Worthypothese gilt dabei als relevant, wenn sie als einzige Worthypothese der Retrievalliste zu mindestens 50% IOU mit einer zum Anfragewort passenden Wortannotation überlappt.
Die mAP ist der Durchschnitt der *Average Precision*s aller Anfragen. Die Average Precision (AP) einer Anfrage ist die Fläche unter der sogenannten *Precision-Recall Kurve* (PR-Kurve), die sich durch die Betrachtung der *Precision* bei bestimmten Recall-Stufen in der Retrievalliste ergibt (siehe [MRS08] Kapitel 8.4). Eine AP von 1 ergibt sich, wenn alle relevanten Exemplare zu einer Anfrage am Anfang der Retrievalliste stehen. Die AP nimmt ab, je mehr irrelevante Exemplare vor relevanten Exemplaren in der Retrievalliste stehen.
Für die Berechnung der mAP haben sich verschiedene Varianten, die zum Teil auf Interpolationen der PR-Kurve basieren, etabliert (siehe [MRS08] Kapitel 8.4). Für den Benchmark zum Kriegstagebuch wird die gleiche Berechnung wie in [RSR+17] verwendet.

Der Benchmark bietet aufgrund der geringen Größe des Datensatzes einige Schwierigkeiten. Beispielsweise kommt der Buchstabe „x“ nur einmal im Kriegstagebuch (im Test) vor. Er kann also nicht beim Training berücksichtigt werden. Der Benchmark enthält außerdem viele kurze Anfragewörter. Kurze Worte können ein Problem für Word Spotting Methoden darstellen [Kas16]. Bei vielen Anfragewörtern handelt es sich um Stoppworte, deren Bedeutung bei der Anwendung von Word Spotting in Frage gestellt werden kann. Beispielsweise sind Stoppworte normalerweise irrelevant für eine Indexierung von Dokumenten. Aufgrund der vergleichsweise geringen Anzahl von Wortannotationen wurden Stoppworte als Anfragewörter in den Benchmark aufgenommen. Eine Schwierigkeit für trainingsbasierte Verfahren ist die vergleichsweise geringe Anzahl an Wörtern in der Trainingsmenge. Die 21 Seiten des TrainVal Teildatensatzes des Kriegstagebuchs enthalten nur 1434 Wörter und damit weniger, als die kleinsten Trainingsmengen (Train I) der Word Spotting Benchmarks der ICFHR2016 (siehe [SF18] Tabelle 1 (Botany und Konzilsprotokolle)).

4.3 ZUSÄTZLICHES TRAININGSMATERIAL

Als Ausgleich für die geringe Menge an Trainingsmaterial im Benchmark zum Kriegstagebuch werden in dieser Arbeit weitere Datensätze historischer Dokumente hinzugezogen. Bei den verwendeten Datensätzen handelt es sich um den Trainings- und Validierungsteil des ICDAR2017 Handwritten Keyword Spotting Competition (ICDAR2017 KWS) Datensatzes[1] (71615 Wortabbilder), die Krupp Briefe (766 Wortab-

1 https://scriptnet.iit.demokritos.gr/competitions/7/ zuletzt abgerufen am 07.04.2018.

bilder) aus [Cha16], die Trainingsdaten III der Konzilsprotokolle (16919 Wortabbilder) und des Botany Datensatzes (21981 Wortabbilder) [2] und um die 4860 Wortabbilder des im Bereich Word Spotting häufig verwendeten George Washington Datensatzes (siehe [SF18] Abschnitt 4.1.1). Die verwendeten Wortannotationen für den ICDAR2017 KWS Datensatz wurden mittels Forced-Alignment wie in [SRF17] beschrieben und auf Grundlage der bereitgestellten Zeilenannotationen erstellt.
Die Auswahl dieser Datensätze erfolgte zum einen, da ein Großteil der Daten frei verfügbar ist, zum anderen enthält die Menge der insgesamt 176594 Wortabbilder viele Bilder deutschsprachiger Worte, die wie die Worte des Kriegstagebuchs in Kurrentschrift geschrieben wurden.

4.4 EXPERIMENTE

Der erste Schritt der vorgestellten Methodik (siehe Kapitel 3) ist das Training des TPP-PHOCNet. Im Rahmen dieser Studienarbeit wurden mehrere Modelle des TPP-PHOCNet trainiert, dabei wurde mit unterschiedlichen Trainingsmengen experimentiert. Die Beurteilung der Qualität der gelernten PHOC-Embeddings zum Word Spotting im Kriegstagebuch erfolgt in Unterabschnitt 4.4.1 zunächst im segmentierungsbasierten Szenario. Grund dafür sind die Vorüberlegungen, die zu Beginn des Unterabschnittes geführt werden.
Der zweite Teil 4.4.2 dieses Abschnittes beschreibt die segmentierungsfreien Experimente, die sich in zwei Phasen unterteilen. In der ersten Phase sind die Experimente darauf ausgelegt hohe Detektionsraten zu erreichen, um eine gute Ausgangslage für das Retrieval zu bieten. In der zweiten Phase wird, die Eignung der Worthypothesen mit den höchsten Detektionsraten zum Word Spotting untersucht.

4.4.1 *Segmentierungsbasiert*

Bevor explizit auf die Experimente eingegangen wird, erfolgt eine kurze Vorüberlegung, inwiefern das segmentierungsbasierte Auswertungsprotokoll mit dem segmentierungsfreien Auswertungsprotokoll, unter Berücksichtigung der verwendeten Methodik, zusammenhängt.

Die vorgegebene Wortsegmentierung im segmentierungsbasierten Szenario entspricht im segmentierungsfreien Vorgehen mit Worthypothesen eben der Menge der Worthypothesen. Aus der segmentierungsfreien Sicht ist das segmentierungsbasierte Szenario also ein Sonderfall mit einer Detektionsrate von 100%. Die Retrievalliste

2 https://www.prhlt.upv.es/contests/icfhr2016-kws/data.html zuletzt abgerufen am 07.04.2018.

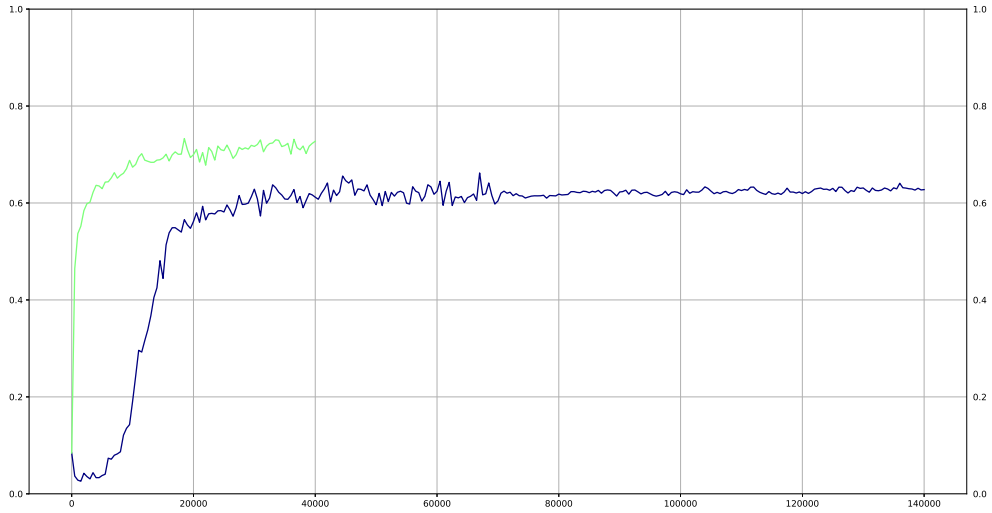


Abbildung 4.4.1: Lernkurven der TPP-PHOCNet Modelle Train (blau) und Bouns+Trian (grün) auf dem Validierungsdatensatz. Dargestellt wird die mAP im Query-by-Example Szenario.

wird nicht gekürzt, daher beträgt auch der mittlere Recall 100%. Würde im segmentierungsfreien Szenario ein IOU-Schwellwert von 100% angenommen werden, dann wäre die vorgegebene Wortsegmentierung in jeder Menge von Worthypothesen enthalten, welche eine Detektionsrate von 100% besäße. Die mAP der Auswertung im segmentierungsbasierten Szenario ist also eine exakte obere Schranke für die mAP im segmentierungsfreien Szenario, bei einem IOU-Schwellwert von 100%. Wegen der Vorüberlegungen lässt sich die segmentierungsbasierte mAP als Indikator für eine obere Schranke der mAP im segmentierungsfreien Szenario ansehen. Das Ziel der segmentierungsbasierten Experimente ist damit das Finden eines TPP-PHOCNet Modells mit möglichst hoher segmentierungsbasierte mAP. Außerdem wird wegen der Ergebnisse in [RSR+17] davon ausgegangen, dass keine weitere Anpassung der im segmentierungsbasierten Szenario trainierten PHOCNet Modelle für das segmentierungsfreie Word Spotting benötigt wird.

Zur Verifizierung, dass die verwendete Implementierung des TPP-PHOCNet lernt, wurde ein Trainingsexperiment mit den 950 Worten des Train Teildatensatzes des Benchmarks durchgeführt. Bei dem Experiment wurde die mAP jeweils nach 500 Iterationsschritten für das Query-by-Example Szenario auf dem Validierungsteil des Benchmarks als Lernkurve aufgezeichnet. Abbildung 4.4.1 zeigt die Lernkurve, die sich aus den aufgezeichneten mAP-Werten ergibt. Zur Identifikation von möglichen

Trainingsmaterial	mAP@QbE	mAP@QbS
Bonus	5.40	10.74
Train	53.17	44.09
Bonus+Train	68.00	78.69
TrainVal	66.54	56.72
Bonus+TrainVal	**72.26**	**82.44**

Tabelle 4.4.1: mAPs der trainierten TPP-PHOCNet Modelle auf dem Testteil des Benchmarks im segmentierungsbasierten Szenario.

Sprüngen in der Lernkurve wurden 140000 Trainingsiterationen durchgeführt, die ersten 70000 bei einer Lernrate von 10^{-4}, danach wurde die Lernrate auf 10^{-5} reduziert. Da sich die mAP nach 80000 Iterationen nicht weiter gesteigert hat wurde beim nächsten Experiment mit dem Standardparameterwerten (siehe Kapitel 3.1) fortgefahren. Die resultierenden TPP-PHOCNet Modelle werden im Folgenden jeweils mit den Namen der (Teil-)Datensätze auf denen sie trainiert wurden bezeichnet. Aus dem ersten Trainingsexperiment ergibt sich damit das *Train* Modell des TPP-PHOCNet.

Das zweite Trainingsexperiment wurde mit den 1434 Wörtern des TrainVal Teildatensatzes des Benchmarks durchgeführt, um eine Qualitätssteigerung durch mehr Trainingsdaten zu überprüfen. Eine Evaluierung des TrainVal Modells auf dem Validierungsteil des Benchmarks ist, wegen der Verwendung der Wortabbilder Validierungsteil beim Training, wenig sinnvoll. Für den Vergleich der Modelle Train und TrainVal wurde die mAP beider Modelle daher auf dem Testdatensatz des Benchmarks ausgewertet. Die mAP des TrainVal Modells ist für beide Anfragetypen auf dem Testdatensatz mit 66.54% (QbE) und 56.72% (QbS) jeweils um etwa 12%-Punkte (absolut) höher als die des Train Modells (siehe Tabelle 4.4.1).

Für die beiden Modelle Train und TrainVal wurden ausschließlich Wortabbilder des Kriegstagebuchs als Trainingsmaterial verwendet. Weitere Trainingsexperimente wurden daher mit Hinblick auf eine Steigerung der mAP durch zusätzliches Trainingsmaterial durchgeführt. Dazu wurde zunächst ein TPP-PHOCNet Modell (*Bonus*) auf den 176594 Wortabbildern des Zusatzmaterials (siehe Abschnitt 4.3) mit den Standardwerten der Lernparameter (siehe Kapitel 3.1) vortrainiert. Anschließend wurden zwei Modelle, jeweils mit dem Train- und dem TrainVal-Teil des Benchmarks für 40000 Iterationen und einer Lernrate von 10^{-5}, wie in [GSF18], nachtrainiert (Bonus+Train bzw. Bonus+TrainVal). Die Lernkurve auf dem Validierungsdatensatz des Bonus+Train Modells ist in Abbildung 4.4.1 dargestellt. Der rapide Anstieg der mAP nach ver-

gleichsweise wenigen Iterationen deckt sich mit den Erkenntnissen aus [GSF18] für andere Datensätze. Bereits nach etwa 2500 Iterationen wird die mAP des Train Modells übertroffen.

Tabelle 4.4.1 zeigt die mAPs aller genannten Modelle, jeweils im Query-by-String (mAP@QbS) und im Query-by-Example (mAP@QbE) Szenario auf dem Testdatensatz. Mit dem Modell ohne Trainingsdaten aus dem Kriegstagebuch (Bonus) wurden erwartungsgemäß keine hohen mAP Werte erreicht. Zusammen mit den Trainingsdaten aus dem Kriegstagebuch wurde eine deutliche Steigerung durch das Pretraining mit dem zusätzlichen Trainingsmaterial erreicht (siehe Tabelle 4.4.1). Das beste Modell, bezüglich der mAP im segmentierungsbasierten Szenario auf dem Testdatensatz, ist das Bonus+TrainVal Modell. Mit diesem Modell wurde eine mAP@QbS von 82.44% und eine mAP@QbE von 72.26% erreicht.

Neben der Menge an Trainingsmaterial bzw. dem Pretraining unterschieden sich die Modelle durch die PHOC Repräsentationen, mit denen sie trainiert wurden. Die PHOC-Repräsentationen wurden bei den Modellen ohne Pretraining jeweils aus den in den Trainingsmengen Train bzw. TrainVal enthaltenen Buchstaben abgeleitet. Bei den Modellen mit Pretraining ergab sich die PHOC-Repräsentation aus den Buchstaben des Zusatzmaterials, wobei auch Sonderzeichen enthalten waren. Im Anhang A.1 sind die jeweiligen Buchstabenmengen der Trainingsdatenmengen angegeben. Um auszuschließen, dass die Unterschiede in der mAP nicht wesentlich mit den unterschiedlichen PHOC-Repräsentationen zusammenhängen, wurden zusätzlich zwei Modelle auf dem Train bzw. TrainVal Teil des Benchmarks trainiert, wobei die gleiche PHOC-Repräsentation wie bei den Bonus Modellen verwendet wurde. Im Vergleich zu den Modellen Train bzw. TrainVal ergab sich keine Verbesserung der mAP auf dem Testdatensatz durch die andere PHOC Repräsentation (siehe Tabelle A.0.1 im Anhang). Damit ist die Steigerung der mAP erwartungsgemäß auf das Pretraining mit zusätzlichem Trainingsmaterial zurückzuführen.

Bevor auf die segmentierungsfreien Experimente eingegangen wird, gibt es eine Auffälligkeit bei den segmentierungsbasierten Experimenten zu bemerken. Die mAP@QbE Werte der Modelle ohne Pretraining sind jeweils höher als die entsprechenden mAP@QbS Werte. Bei den Modellen mit Pretraining ist es genau umgekehrt. Als Grund dafür wird die Menge an Trainingsmaterial gesehen. Je mehr Trainingsmaterial verwendet wird, desto besser lernt das PHOCNet das eigentliche Ziel, die Vorhersage von PHOC-Vektoren, welche den binären PHOC-Vektoren von Strings gleichen.

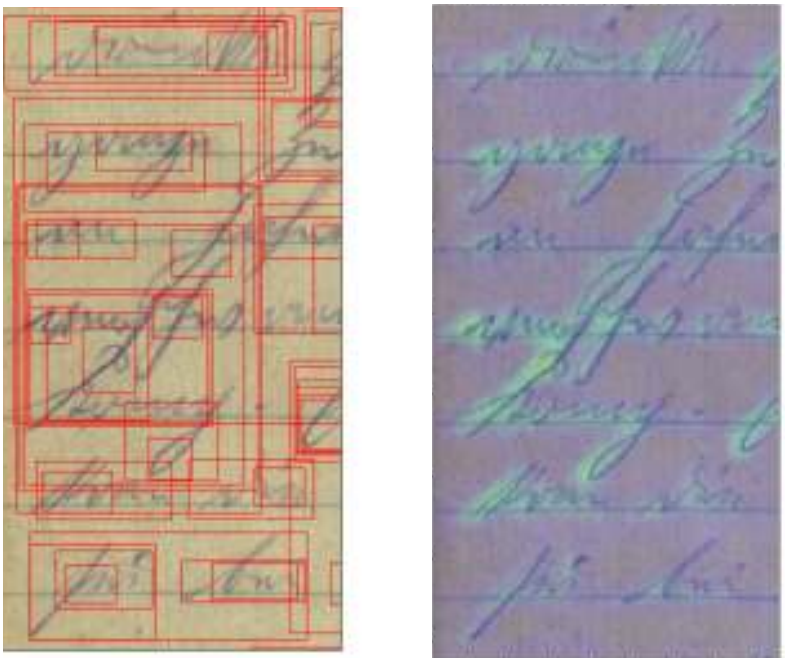
4.4.2 *Segmentierungsfrei*

Ein limitierender Faktor der verwendeten segmentierungsfreien Word Spotting Methodik sind die Worthypothesen. Nicht detektierte Wörter sind zwangsläufig nicht in der Retrievalliste enthalten. Somit begrenzt die Detektionsrate den mittleren Recall und dieser begrenzt wiederum die mittlere Average Precision. Die ersten Experimente beschränken sich daher auf die Generierung der Worthypothesen, mit dem Ziel, die Detektionsrate zu erhöhen. Anschließend erfolgt eine abschließende Auswertung des segmentierungsfreien Word Spottings durch die Kombination des besten TPP-PHOCNet Modells (Bonus+TrainVal) und den besten Worthypothesen auf dem Testdatensatz.

Worthypothesen

Es wurden Experimente zu den Hypothesen der in Kapitel 3.2 angesprochenen Textdetektoren durchgeführt. Sofern nicht anders angegeben, wurden bei den Experimenten die Standardparameterwerte, wie in [RSR+17] bzw. wie in Kapitel 3.2 angegeben, verwendet.

SIFT-KONTRAST TEXTDETEKTOR Die Standardeinstellung des SIFT-Kontrast Textdetektors aus [RSR+17] lieferte auf dem Validierungsteil des Benchmarks lediglich eine Detektionsrate von 36.00%, wobei durchschnittlich 4596 Hypothesen pro Seite erzeugt wurden. In informellen Validierungsexperimenten zum Sift-Kontrast Textdetektor wurde daher eine Parameterkonfiguration des SIFT Deskriptor gesucht, mit der eine bessere Detektionsrate erreicht wird. Mit einer Zellengröße von 8 Pixeln, einer Zellenstruktur von 4×4 Zellen und einem Deskriptorabstand von 4 Pixeln wurde eine Detektionsrate von 67.90% auf dem Validierungsdatensatz erreicht. Dabei wurden durchschnittlich 679 Hypothesen pro Seite erzeugt. Die ermittelten Parameterwerte des SIFT-Kontrast Textdetektors wurden für alle weiteren Experimente festgehalten.
Bei den erwähnten Validierungsexperimenten wurden 50 ER-Schwellwerte verwendet. Durch eine Erhöhung der Anzahl der ER-Schwellwerte auf 250 konnte die Detektionsrate mit dem SIFT-Kontrast Textdetektor auf 75.30% auf dem Validierungsdatensatz gesteigert werden. Die Anzahl der durchschnittlichen Worthypothesen pro Seite hat sich entsprechend auf 1446 erhöht. Im Vergleich zur Hypothesenanzahl in [RSR+17] von etwa 10000 Hypothesen pro Seite ist die Anzahl von 1446 Hypothesen, zumindest bezüglich Laufzeitüberlegungen, eher gering.
Eine qualitative Betrachtung der SIFT-Kontrast Werte und der daraus erzeugten ER Worthypothesen zeigte einige Schwächen der Heuristik. Kontrastvariationen der



(a) Worthypothesen des SIFT-Kontrast Textdeketors bei 50 ER-Schwellwerten.

(b) Grundlage der Wort Hyopthesen, sind die SIFT-Kontrast Werte.

Abbildung 4.4.2: Beispiel der Worthypothesen des SIFT-Kontrast Textdetektors. Zur besseren Ansicht sind nur Worthypothesen mit einer Fläche von mehr als 2000 Pixeln dargestellt.

Schrift, insbesondere Kontrastvariationen innerhalb einzelner Wörter, stellen ein Problem im Kriegstagebuch dar. Durch die Zeilenlinien sind die Wörter einer Zeile in horizontaler Richtung durch vergleichsweise kontrastreiche Bildregionen miteinander verbunden. Die Annahme der ER-Heuristik (vgl. [RSR+17] Abschnitt III.C), dass lokale Minima der Textscores typischerweise Wortzwischenräumen entsprechen, trifft für die SIFT-Kontrast Werte als Textscores im Kriegstagebuch nicht immer zu. Ein gravierenderes Problem ist jedoch, dass Ober- und Unterlängen benachbarter Zeilen teilweise ineinander übergehen. Wörter sind daher nicht an allen Zeilen durch Wortzwischenräume im Schriftbild voneinander getrennt. Abbildung 4.4.2 zeigt ein Beispiel, bei dem der Kontrast an einem Übergang so groß ist, sodass sich keine zwei Worthypothesen an diesem Übergang bilden. Stattdessen werden die Worthypothesen erzeugt, welche diesen Übergang umschließen. In der Abbildung sind auch solche

Worthypothesen zu sehen, welche die Schrift einzelner Wörter akkurat erfassen.
Im Vergleich zu den Kontrastvariationen der Schrift wird der generell auffällig schwache Kontrast der Schrift in den Dokumentenabbildern (siehe Abbildung 4.4.2) als geringeres Problem der Heuristik gesehen, das durch eine genügend große Anzahl an ER-Schwellwerten umgangen werden kann.

Mit Hinblick auf eine spätere Kombination verschiedener Textdetektoren werden die Experimente zu AAM-basierten Worthypothesen im nächsten Paragraphen erläutert.

AAM TEXTDETEKTOR Am Anfang dieser Experimente zum AAM Textdetektor stand die Vermutung, dass eine größere Menge an Trainingsmaterial zu einer Steigerung der Detektionsrate führt, ähnlich wie bei den segmentierungsbasierten Experimenten zum TPP-PHOCNet. Zur Überprüfung dieser These wurden mehrere AAM-PHOCNet Modelle entsprechend der Trainingsmengen, welche auch in den segmentierungsbasierten Experimenten zum TPP-PHOCNet eingesetzt wurden, trainiert. Das Training der AAM-PHOCNet Modelle Bonus, Train und TrainVal wurde dabei jeweils für 80000 Iterationen und mit den Standardparameterwerten (siehe Kapitel 3.1) durchgeführt. Die AAM-PHOCNet Modelle Bonus+Train und Bonus+TrainVal sind durch Nachtrainieren des Bonus Modells für 40000 Iterationen bei einer Lernrate von 10^{-5} hervorgegangen. Diese Modelle wurden anschließend für den AAM Textdetektor mit Quantisierung (siehe Kapitel 3.2) verwendet und entsprechende Detektionsraten sowie die durchschnittliche Anzahl von Worthypothesen je Seite, auf dem Validierungsdatensatz ermittelt. Die Tabelle 4.4.2 zeigt die Ergebnisse, bei einer Einstellung von 50 ER-Schwellwerten.

Worthypothesen (aam-quant)	Detektionsrate	#Worthypothesen
Bonus	29.21	1416
Train	50.20	1076
Bonus+Train	73.66	1279
TrainVal	28.39	1210
Bonus+TrainVal	66.04	1352

Tabelle 4.4.2: Detektionsraten (in%) und durchschnittliche Anzahl von Worthypothesen je Seite der verschiedenen AAM Textdetektoren auf dem Validierungsdatensatz des Benchmarks.

Worthypothesen	Detektionsrate	#Worthypothesen
sift-quant (sq) (50)	67.90	679
aam-quant (aq) (50)	73.66	1279
aam-sift-quant (asq) (50)	70.57	955
sq+aq (50)	88.47	1958
sq+aq+asq (50)	90.32	2913
sq+aq+asq (250)	94.44	7963

Tabelle 4.4.3: Detektionsraten (in %) und durchschnittliche Anzahl von Worthypothesen je Seite verschiedener Kombinationen von Worthypothesen und Textdetektoren. Die dargestellten Auswertungsergebnisse wurden auf dem Validierungsdatensatz ermittelt.

Die Experimente zeigen einen möglichen positiven Einfluss des Pretrainings mit dem zusätzlichen Trainingsmaterial (siehe Tabelle 4.4.2). Die größere Trainingsmenge TrainVal hat im Vergleich zur Train zu geringeren Detektionsraten geführt.

In den folgenden Experimenten zu Kombinationen der Textdetektoren und Worthypothesen wird, sofern nicht anders angegeben, das Bonus+Train Modell des AAM-PHOCNet als Basis für den AAM Textdetektor verwendet, da mit diesem Modell die höchsten Detektionsraten mit 73.66% auf dem Validierungsdatensatz erreicht werden.

KOMBINATIONEN Zur Steigerung der Detektionsrate wurden Experimente zu Kombinationen auf verschiedenen Ebenen der Worthypothesen Methodik durchgeführt. In [RSR+17] wird demonstriert, dass eine Linear-Kombination verschiedener Textdetektoren zu einer Menge von Worthypothesen mit höherer Detektionsrate führen kann. Experimente zur Linear-Kombination der SIFT-Kontrast Textscores und der AAM-Textscores (aam-sift-quant (asq)) bestätigen dieses Ergebnis für den Validierungsdatensatz des Kriegstagebuchs zwar nicht (siehe Tabelle 4.4.3), allerdings kann die Detektionsrate durch eine andere naheliegende Kombinationsmöglichkeit, nämlich die Vereinigung der Worthypothesen verschiedener Textdetektoren gesteigert werden.

Die Experimente zu solchen vereinigten Mengen von Worthypothesen (sq+aq) zeigen, dass der SIFT-Kontrast Textdetektor (sift-quant (sq)) und der AAM Textdetektor (aam-quant (aq)) unterschiedliche Worthypothesen erzeugen. Durch die Kombination der Worthypothesen beider Textdetektoren wird eine Detektionsrate von 88.47% auf dem Validierungsdatensatz erreicht. Dabei enthält die Vereinigung durchschnittlich 1958 Worthypothesen pro Seite des Validierungs- bzw. Testdatensatzes (siehe Tabelle

4.4.3).
Durch weiteres Hinzufügen der Worthypothesen der Linear-Kombination (aam-sift-quant (asq)) kann die Detektionsrate auf 90.32% gesteigert werden (sq+aq+asq). Die durchschnittliche Anzahl der Worthypothesen je Seite steigt dabei auf 2913 (siehe Tabelle 4.4.3). Die meisten Worthypothesen, welche durch die Linear-Kombination erzeugt werden, unterscheiden sich somit ebenfalls von denen der beiden Einzelkomponenten der Linear-Kombination.
Die mit (50) gekennzeichneten Einträge zu Worthypothesen in Tabelle 4.4.3 wurden jeweils bei 50 ER-Schwellwerten ermittelt. Eine Erhöhung dieser Anzahl der ER-Schwellwerte auf 250 führte bei der Vereinigung der Worthypothesen sq+aq+asq (250) zu einer maximalen Detektionsrate von 94.44% bei durchschnittlich 7963 Worthypothesen je Seite des Validierungsdatensatzes.

Die beiden Mengen von Worthypothesen mit den höchsten Detektionsraten (sq+aq+asq (50)) und (sq+aq+asq (250)) wurden für die folgenden Word Spotting Experimente genutzt.

Word Spotting

Die folgenden Experimente dienen als abschließende Auswertung der Word Spotting Methodik im Kriegstagebuch. Tabelle 4.4.4 zeigt die Detektionsrate, den mittleren Recall und mittlere Average Precision der segmentierungsfreien Word Spotting Methodik für drei Kombinationen von Worthypothesen und TPP-PHOCNet Modellen. Die ersten beiden Zeilen verwenden das Bouns+TrainVal TPP-PHOCNet zum Retrieval (siehe Abschnitt 4.4.1) und die im letzten Abschnitt genannten Mengen von Worthypothesen. Die letzte Zeile (sq+aq+asq* (250)) der Tabelle dient als Referenz zum Fall ohne zusätzliches Trainingsmaterial. Dabei wurde der AAM Textdetektor mit dem Train AAM-PHOCNet Modell für die AAM Worthypothesen verwendet und das TrainVal TPP-PHOCNet Modell (siehe Tabelle 4.4.1) für das Retrieval. Die höchste mAP im Query-by-Example Szenario beträgt 58.97% (siehe Tabelle 4.4.4). Die höchste mAP im Query-by-String Szenario beträgt nur 52.29%, obwohl das verwendete TPP-PHOCNet Modell eine höhere mAP@QbS im segmentierungsbasierten Fall erreicht hat (siehe Abschnitt 4.4.1). Grund dafür ist der Benchmark. Im Query-by-Example Szenario ist das Anfragewort selbst bei der Auswertung enthalten. Sobald das Query-by-Example Anfragewort präzise von einer Wort Hypothese erfasst, wird das Retrievalproblem des selbigen trivial.

In beiden Experimenten fällt auf, dass der Abfall von Detektionsrate zu mittlerem Recall und vom mittleren Recall zu mittlerer Average Precision relativ hoch ist (siehe Tabelle 4.4.4). Eine Auswertung des Experimentes (sq+aq+asq (250)) im QbS Szenario

	QbE			QbS		
Worthypothesen	DR	mR	mAP	DR	mR	mAP
sq+aq+asq (50)	86.53	72.47	55.55	86.53	71.62	51.33
sq+aq+asq (250)	90.79	75.21	**58.97**	90.79	75.20	**52.29**
sq+aq+asq* (250)	89.01	71.89	49.96	89.01	64.23	32.87

Tabelle 4.4.4: Query-by-Example (QbE) und Query-by-String (QbS) Evaluierungsergebnisse im segmentierungsfreien Word Spotting Szenario auf den Testdaten des Benchmarks. Angegeben sind die Detektionsrate (DR), mittlerer Recall (mR) und die mittlere Average Precision (mAP) jeweils in %.

bei einem IOU-Schwellwert von 30% hat eine DR von 99.72%, einen mR von 92.56% und eine mAP 59.86% ergeben. Der Unterschied zwischen der Detektionsrate und dem mittleren Recall, der sich in Tabelle 4.4.4 zeigt, ist für eine qualitative Anwendung, wie z.B. das Durchsuchen des Kriegstagebuchs, daher nicht gravierend. Der Unterschied zwischen mR und mAP ist allerdings auch bei einem IOU-Schwellwert von 30% relativ hoch.
Ein Grund dafür sind Schwierigkeiten beim Retrieval kurzer Anfragewörter. Im Query-by-String Szenario liegt die mAP der Anfragewörter mit einem Buchstaben bei weniger als 4% bei beiden gezeigten Experimenten aus Tabelle 4.4.4. Die Abbildung A.0.1 im Anhang zeigt ein Balkendiagramm, welches die mAP-Werte für verschiedene Anfragewortlängen zeigt. Eine qualitative Betrachtung von Retrievallisten (siehe Abbildung 4.4.3) zeigt außerdem, dass für solche kleinen Wörter teilweise sehr unplausible Worthypothesen z.B. vom Einband oder von den Rändern der Dokumentenseiten hoch bewertet werden. In anderen Fällen sind die kleinen Anfragewörter in längeren Wörtern enthalten.
Ein weiterer Grund sind niedrige mAP Werte bei langen Anfragewörtern, da diese nicht immer akkurat durch Worthypothesen erfasst werden. Dabei lässt sich auch beobachten, dass Worthypothesen, welche mehrere Wörter enthalten, weit oben in der Retrievalliste stehen (siehe Anhang 4.4.3).

Eine Zusammenfassung der wichtigsten Evaluierungserkenntnisse erfolgt im nächsten Kapitel.

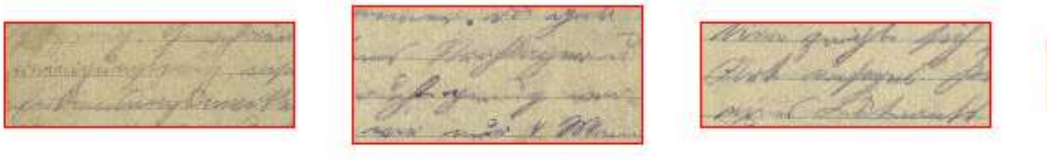
(a) Top 8 Retreival Ergebnisse der QbS Anfrage „2“ enthält keinen einzigen Treffer, stattdessen sind Wort Hypothesen vom Einband des Kriegstagebuchs sowie Seitenränder enthalten.

(b) Top 8 Retreival Ergebnisse der QbS Anfrage „4“ enthält einen Treffer.

(c) Top 15 Retreival Ergebnisse der QbE Anfrage „in“ enthält 5 Treffer. Eine visuelle Ähnlichkeit zu den nicht Treffern ist erkennbar. Einige der nicht Treffer stammen aus größeren Wörten.

(d) Top 5 Retreival Ergebnisse der QbE Anfrage „nachmittags“ enthält 4 von 4 möglichen Treffer. Eine visuelle Ähnlichkeit zu den nicht Treffern ist erkennbar. Einige der nicht Treffer stammen aus größeren Wörten.

(e) Top 6 Retreival Ergebnisse der QbS Anfrage „abends“ enthält 5 Treffer.



(f) Top 3 Retreival Ergebnisse der QbS Anfrage „erkennungsmarke“. Die erste Wort Hypothese enthält das Gesuchte Wort, wird allerding nicht als Treffer gezählt.

Abbildung 4.4.3: Qualitative Beispiele zum segmentierungs-freien Word Spotting mit den „besten“ Wort Hypothesen und dem „besten“ TPP-PHOCNet Modell (siehe 4.4.2). Das Retrieval wurde im Testdatensatz durchgeführt. Gezeigt werden jeweils die ersten X Einträge der Retrievallisten verschiedener Query-by-String und Query-by-Example Anfrageworte. Im Query-by-Example Fall ist das erste gezeigte Wort mit Schwarzer Umrandung das Anfragewort. Relevante Worthypothesen sind grün umrandet, irrelevante Worthypothesen sind rot umrandet.

5 FAZIT UND AUSBLICK

In diesem abschließenden Kapitel wird zunächst eine Übersicht über die Kapitel der Studienarbeit gegeben und deren wichtigsten Punkte herausgestellt. Anschließend erfolgt eine Zusammenfassung der Evaluierungsergebnisse, wobei auch Ideen für zukünftige Experimente als Ausblick gegeben werden. Abschließend werden die erzielten Word Spotting Ergebnisse anwendungsorientierte beurteilt.

Die vorliegende Studienarbeit hat sich mit CNN-basiertem Word Spotting in einem historischen Datensatz, dem Kriegstagebuch (siehe Kapitel 1.1) befasst. Der erste Beitrag dieser Arbeit ist die Vorbereitung des Kriegstagebuchs für die Evaluierung der Methodik. Dazu wurden Bounding Box Wortannotationen erstellt und ein Benchmark festgelegt (siehe Kapitel 4.1 und 4.2).
Die Auswahl der Word Spotting Methodik erfolgte zum einen wegen der State-of-the-Art Ergebnisse CNN-basierter Verfahren und zum anderen wegen der Möglichkeit eines Ausgleichs kleiner Trainingsmengen durch Pretraining auf zusätzlichem Trainingsmaterial, das von anderen Datensätzen stammt oder synthetisch generiert wird (siehe Kapitel 2 Teil 1). Für die ausgewählte CNN-basierte Methodik sowie für andere CNN-basierte Methoden wurde festgestellt, dass sie sich wegen der verwendeten Attributrepräsentationen, wie z.B. der PHOC Repräsentation, so erfolgreich zum Word Spotting einsetzen lassen (siehe Kapitel 2 Teil 2). Dabei wurde auch festgestellt, dass der Einsatz CNN-basierter Verfahren zum segmentierungsfreien Word Spotting derzeit meist auf einer Vorberechnung von Attributrepräsentationen für Worthypothesen erfolgt (siehe Kapitel 2 Teil 2). In Kapitel 3 wurde die Methodik mit Hinblick auf wichtige Ideen und die Nachvollziehbarkeit der Evaluierung skizziert.

Der zweite Beitrag der vorliegenden Arbeit sind die Experimente zur Evaluierung zur ausgewählten Methodik im Kriegstagebuch (siehe Kapitel 4.4). Eine erste Feststellung ist, dass der Datensatz und der Benchmark Schwierigkeiten aufweisen. Zu den Schwierigkeiten des Datensatzes gehören die Überschneidungen der Schrift verschiedener Zeilen und Kontrastvariationen. Trotz dieser Schwierigkeiten wurden eine hohe Detektionsrate über 90% mit dem Worthypothesen Ansatz erreicht. Entscheidend dafür war die Vereinigung verschiedener Mengen von Worthypothesen. Es wurde festgestellt, dass sich die Mengen der Worthypothesen verschiedener Textdetektoren unterscheiden. Die Steigerung der Detektionsrate wirkt sich auch positiv auf die

segmentierungsfreie Word Spotting Qualität (gemessen als mAP) aus (siehe Tabelle 4.4.4), obwohl mit ihr eine deutliche Erhöhung der Anzahl an Worthypothesen und damit eine Erschwerung der Auswahl und des Sortierung relevanter Worthypothesen einhergeht.
Zu den Schwierigkeiten des Benchmarks zum Kriegstagebuch gehören die geringe Menge an Trainingsmaterial und das vorhanden sein kurzer Anfragewörter. Durch den Einsatz des Pretrainings auf zusätzlichem Trainingsmaterial wurde, ähnlich zu [GSF18], demonstriert, dass der ersten Schwierigkeit auch im Kriegstagebuch entgegengewirkt werden kann. Ein Vergleich der Ergebnisse ohne zusätzliches Trainingsmaterial mit den Ergebnissen mit zusätzlichem Trainingsmaterial zeigt eine deutliche Steigerung der mAP-Werte (siehe Tabelle 4.4.1 und 4.4.4).

Bei der Schwierigkeit der kurzen Anfragewörter war auffällig, dass unplausible Worthypothesen weit vorne in der in der Retrieval Liste stehen. Als Idee zur Vermeidung dieses Problems wird für zukünftige Experimente vorgeschlagen negativ Beispiele, also Bilder aus den Dokumenten, welche keine Wörter enthalten, mit in das Training des TPP-PHOCNets einzubeziehen. Denkbar wäre dazu die Einführung eines oder mehrerer zusätzlichen Attribute, die beschreiben ob und um was für eine Art von negativ Beispiel es sich handelt.

Insgesamt lässt sich sagen, dass mit der segmentierungsfreien Word Spotting Methodik Ergebnisse erzielt wurden, die eine Anwendung zur gezielten Suche nach Wörtern im Kriegstagebuch ermöglichen. Die meisten Vorkommen gesuchter Wörter können gefunden werden, dass zeigen die hohen Detektionsraten von über 90%. Von diesen Vorkommen sind die meisten auch in der Retreivalliste enthalten, dass zeigt der hohe mittlere Recall, der bei einem IOU-Schwellwert von 30% ebenfalls über 90% liegt (siehe Kapitel 4.4.2). Die mAP von 58.97% im Query-by-Example und 52.29% im Query-by-String Szenario zeigen zudem, dass relevante Worthypothesen zu einer Anfrage normalerweise weit vorne in der Retrievalliste stehen und nur mit wenigen irrelevanten Worthypothesen vermischt sind. Abschließend sei angemerkt, dass durch die Vorberechnung der PHOC-Repräsentationen (siehe Kapitel 3) für die durchschnittlich etwa 7900 Worthypothesen je Seite (siehe Tabelle 4.4.3) Retrievalzeiten von durchschnittlich 96 Millisekunden je Seite erreicht wurden. Zusammen mit einer möglichen Parallelisierung des Retrievals über mehrere Seiten kann so extrem schnelle Suche (im Kriegstagebuch) realisiert werden.

A ANHANG

Trainingsmaterial	mAP@QbE	mAP@QbS
Train	53.17	44.09
Train (Bonus Unigramme)	49.92	39.64
TrainVal	66.54	56.72
TrainVal (Bonus Unigramme)	65.24	56.43

Tabelle A.0.1: Mittlere Average Precisions der TPP-PHOCNet Modelle, welche allein auf den Daten des Kriegstagebuchs trainiert wurden, auf dem Testdatensatz des Benchmarks (in %). Die PHOC Repräsentation des Train und des TrainVal Modells ergibt sich jeweils aus den Buchstaben der entsprechenden Trainingsmengen (siehe A.1). Bei den Modellen mit dem Zusatz (Bonus Unigramme) wurden die 66 Unigramme des zusätzlichen Trainingsmaterials verwendet. Tendenziell führt die größere Repräsentation mit mehr Unigrammen zu geringeren mAP Werten.

A.1 PHOC UNIGRAMME

- Train (39): -, 0, 1, 2, 3, 4, 6, 7, 8, 9, a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, z, ß, ä, ö, ü
- TrainVal (41): -, /, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, z, ß, ä, ö, ü
- Bonus (66): ", &, ’, (,), *, +, „ -, ., /, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, :, ;, =, ?, @, [,], _, a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z, £, §, ¬, ß, à, ä, é, ö, ü, ˆ, |

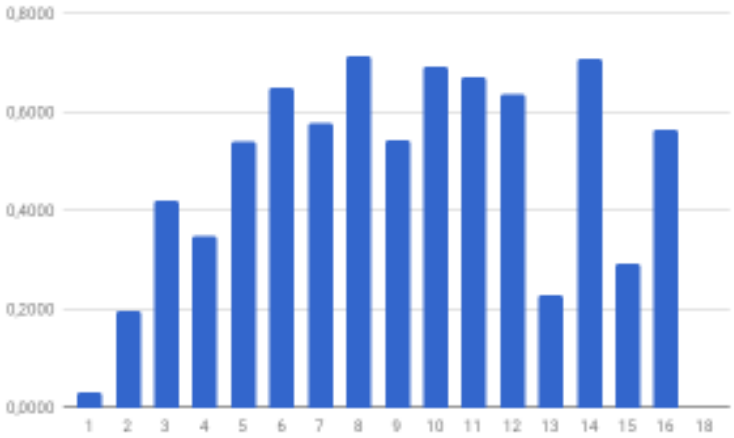


Abbildung A.0.1: Mittlere Average Precisions für verschiedene Längen von Anfragewörtern im segmentierungsfreien QbS Szenario auf dem Testdatensatz des Benchmarks zum Kriegstagebuch. Die Zahlen beziehen sich auf das Experiment sq+aq+asq (250) (siehe Tabelle 4.4.4). Die mAP-Werte kurzer Anfragewörter bis zu einer Länge von 4 Zeichen liegen deutlich unter der mAP von 52.29% (siehe Tabelle 4.4.4). Auch die mAP-Werte der Anfragewörter mit 13, 15 und 18 Zeichen liegen unterhalb dieser Grenze.

LITERATURVERZEICHNIS

[AGFV14] Almazán, J. ; Gordo, A. ; Fornés, A. ; Valveny, E.: Word Spotting and Recognition with Embedded Attributes. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36 (2014), Dec, Nr. 12, S. 2552–2566

[Cha16] Chai, Keng H.: Textdetektion und Indizierung in historischen Dokumenten (Bachelorarbeit). In: *Interne Berichte*, Technische Universität Dortmund, 2016. – Veröffentlicht unter http://patrec.cs.tu-dortmund.de/pubs/theses/ba_chai.pdf (zuletzt Abgerufen am 07.04.2018).

[GBC16] Goodfellow, Ian ; Bengio, Yoshua ; Courville, Aaron: *Deep Learning*. MIT Press, 2016. – http://www.deeplearningbook.org

[GSF18] Gurjar, Neha ; Sudholt, Sebastian ; Fink, Gernot A.: Learning Deep Representations for Word Spotting Under Weak Supervision. In: *Int. Workshop on Document Analysis Systems*, 2018

[GSGN17] Giotis, Angelos P. ; Sfikas, Giorgos ; Gatos, Basilis ; Nikou, Christophoros: A survey of document image word spotting techniques. In: *Pattern Recognition* 68 (2017), S. 310 – 332. – ISSN 0031–3203

[GV17] Ghosh, Suman ; Valveny, Ernest: R-PHOC: Segmentation-Free Word Spotting using CNN. In: *CoRR* abs/1707.01294 (2017). http://arxiv.org/abs/1707.01294

[HZRS14] He, Kaiming ; Zhang, Xiangyu ; Ren, Shaoqing ; Sun, Jian: Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. In: *Computer Vision – ECCV 2014*. Cham : Springer International Publishing, 2014, S. 346–361

[JSD+14] Jia, Yangqing ; Shelhamer, Evan ; Donahue, Jeff ; Karayev, Sergey ; Long, Jonathan ; Girshick, Ross ; Guadarrama, Sergio ; Darrell, Trevor: Caffe: Convolutional Architecture for Fast Feature Embedding. In: *Proceedings of the 22Nd ACM International Conference on Multimedia*. New York, NY, USA : ACM, 2014 (MM '14), 675–678

[Kas16] Kasperidus, Matthias J.: Histogramm-basierte Merkmale zum Word Spotting auf historischen Dokumenten (Bachelorarbeit). In: *Interne Berichte*, Technische Universität Dortmund, Fakultät für Informatik, 2016. – Veröffentlicht unter http://patrec.cs.tu-dortmund.de/pubs/theses/ba_kasperidus.pdf (zuletzt Abgerufen am 07.04.2018).

[Low04] Lowe, David G.: Distinctive Image Features from Scale-Invariant Keypoints. In: *International Journal of Computer Vision* 60 (2004), S. 91–110

[LSD15] Long, J. ; Shelhamer, E. ; Darrell, T.: Fully convolutional networks for semantic segmentation. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. – ISSN 1063–6919, S. 3431–3440

[MCUP04] Matas, J ; Chum, O ; Urban, M ; Pajdla, T: Robust wide-baseline stereo from maximally stable extremal regions. In: *Image and Vision Computing* 22 (2004), Nr. 10, S. 761 – 767. – ISSN 0262–8856

[MHR96] Manmatha, R. ; Han, Chengfeng ; Riseman, E. M.: Word spotting: a new approach to indexing handwriting. In: *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1996. – ISSN 1063–6919, S. 631–637

[MRS08] Manning, Christopher D. ; Raghavan, Prabhakar ; Schütze, Hinrich: *Introduction to Information Retrieval*. New York, NY, USA : Cambridge University Press, 2008. – ISBN 0521865719, 9780521865715

[NB17] Nils Brinkmann, Felix Gonsior Simon Schröder Stefan Hesse Dominik Schütgens Matthias Kasperidus Patrick Trockel Weihan Pang Emily Veuhoff Damian P. Christian Pionzewski P. Christian Pionzewski: PG 602: Erklär mir die Welt – Kamerabasierte Internetrecherche (Abschlussbericht). In: *Interne Berichte*, Technische Universität Dortmund, Fakultät für Informatik, Lehrstuhl 12, 2017

[RATL11] Rusinol, M. ; Aldavert, D. ; Toledo, R. ; Llados, J.: Browsing Heterogeneous Document Collections by a Segmentation-Free Word Spotting Method. In: *2011 International Conference on Document Analysis and Recognition*, 2011. – ISSN 1520–5363, S. 63–67

[RF16] Rothacker, Leonard ; Fink, Gernot A.: Robust Output Modeling in Bag-of-Features HMMs for Handwriting Recognition. In: *Proc. Int. Conf. on Frontiers in Handwriting Recognition*. Shenzhen, China, 2016

[RRF13] Rothacker, Leonard ; Rusinol, Marcal ; Fink, Gernot A.: Bag-of-Features HMMs for Segmentation-Free Word Spotting in Handwritten Documents. In: *Proc. Int. Conf. on Document Analysis and Recognition*. Washington DC, USA, 2013

[RRRG89] Rohlicek, J. R. ; Russell, W. ; Roukos, S. ; Gish, H.: Continuous hidden Markov modeling for speaker-independent word spotting. In: *International Conference on Acoustics, Speech, and Signal Processing,*, 1989. – ISSN 1520–6149, S. 627–630 vol.1

[RSR+17] Rothacker, Leonard ; Sudholt, Sebastian ; Rusakov, Eugen ; Kasperidus, Matthias ; Fink, Gernot A.: Word Hypotheses for Segmentation-free Word Spotting in Historic Document Images. In: *Proc. Int. Conf. on Document Analysis and Recognition*. Kyoto, Japan, 2017

[SF16] Sudholt, Sebastian ; Fink, Gernot A.: PHOCNet: A Deep Convolutional Neural Network for Word Spotting in Handwritten Documents. In: *Proc. Int. Conf. on Frontiers in Handwriting Recognition*. Shenzhen, China, 2016

[SF17] Sudholt, Sebastian ; Fink, Gernot A.: Evaluating Word String Embeddings and Loss Functions for CNN-based Word Spotting. In: *Proc. Int. Conf. on Document Analysis and Recognition*. Kyoto, Japan, 2017

[SF18] Sudholt, Sebastian ; Fink, Gernot A.: Attribute CNNs for Word Spotting in Handwritten Documents. In: *Int. Journal on Document Analysis and Recognition* (2018)

[SRF17] Sudholt, Sebastian ; Rothacker, Leonard ; Fink, Gernot A.: Query-by-Online Word Spotting Revisited: Using CNNs for Cross-Domain Retrieval. In: *Proc. Int. Conf. on Document Analysis and Recognition*. Kyoto, Japan, 2017

[WB16] Wilkinson, T. ; Brun, A.: Semantic and Verbatim Word Spotting Using Deep Neural Networks. In: *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2016. – ISSN 2167–6445, S. 307–312

[WLB17] Wilkinson, Tomas ; Lindström, Jonas ; Brun, Anders: Neural Ctrl-F: Segmentation-free Query-by-String Word Spotting in Handwritten Manuscript Collections. In: *CoRR* abs/1703.07645 (2017). http://arxiv.org/abs/1703.07645

[ZKL+16] Zhou, B. ; Khosla, A. ; Lapedriza, A. ; Oliva, A. ; Torralba, A.: Learning Deep Features for Discriminative Localization. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, S. 2921–2929