technische universität
dortmund

Histogramm-basierte Merkmale zum Word Spotting

auf historischen Dokumenten

Matthias Juri Kasperidus
16. Juni 2016

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

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

INHALTSVERZEICHNIS

1 EINLEITUNG

Im Gegensatz zur Texterkennung in maschinengeschriebenen Dokumenten, die weitgehend als gelöstes Problem der Informatik gesehen werden kann, ist die Texterkennung in handgeschriebenen Dokumenten für viele Anwendungsbereiche bisher nicht adäquat gelöst worden (siehe [Fin14] Kapitel 2.2).
Für maschinengeschriebene Dokumente existieren sogenannte Optical Character Recognition (OCR) Techniken, um eine automatische Transkription von Dokumentenabbildern durchzuführen. Dabei werden die einzelnen Buchstaben im Dokument nach vorheriger Strukturanalyse segmentiert und nachfolgend im Wort- und Textkontext erkannt. Dieses Vorgehen ist unter anderem wegen der hohen Genauigkeit heutiger Drucktechniken und einheitlichen Schriftarten möglich. (siehe [Fin14] Kapitel 2.2).
Das ändert sich allerdings, wenn man Handschriften betrachtet. Häufig ist die Schrift kursiv und von Schreiber zu Schreiber unterschiedlich. Aber auch die Variation gleicher Worte des gleichen Schreibers kann bereits sehr hoch sein. Abbildung 1.0.1 zeigt ein Beispiel, bei dem die Form des Buchstaben ”f” im Wort ”of” bereits sehr stark variiert, obwohl das Wort an drei verschiedenen Stellen desselben Dokumentes ausgeschnitten wurde. Erkennungssysteme für Handschriften müssen also mindestens mit den Schriftvariationen eines Schreibers, auch intra-writer Variabilität [RP08] genannt, umgehen können. Stammen die Handschriften von mehreren Schreibern, so muss ein solches System zusätzlich mit der sogenannten inter-writer Variabilität [RP08] umgehen können.
Bei historischen Handschriften kommen zusätzliche Schwierigkeiten hinzu, so können beispielsweise Alterungsprozesse und andere äußere Einflüsse die Qualität solcher Dokumente beeinträchtigt haben. Außerdem gibt es vergangene Schrift- und Schreibstile, die eine korrekte Texterkennung zusätzlich erschweren [FRG14].
Für historische handschriftliche Dokumente gibt es zur Zeit kaum, wenn nicht keine, brauchbaren OCR Techniken, die gute Ergebnisse erzielen [RM07]. Projekte, wie das "Transcriptorium-Projekt"[1] der EU, zeigen aber, dass großes Interesse an computergestützten Methoden besteht, um Wissen aus großen Datensätzen historischer Dokumente effizient erschließbar zu machen. Eine vielversprechende Technik, um dieses Bedürfnis zu erfüllen ist **Word Spotting**. Beim Word Spotting wird eine kom-

1 http://transcriptorium.eu/

plette Texterkennung umgangen und damit auch viele Schwierigkeiten im Kontext historischer Handschriften [RM07],[MHR96].

1.1 WORD SPOTTING

Word Spotting bezeichnet die Aufgabe, die Vorkommen eines Anfragewortes innerhalb einer Menge von Dokumentenabbildern zu finden [MHR96].

Word Spotting lässt sich somit als *Bildretrieval* auffassen. Es soll eine *Antwortliste* mit Bildausschnitten der Dokumente zurückgegeben werden, in denen das gegebene Anfragewort steht. Je nach Anwendung können sich verschiedene Anforderungen an diese Antwortliste ergeben. Eine naheliegende Anforderung ist die *Vollständigkeit*, d.h. es sollten alle Bildausschnitte, die der Anfrage entsprechen, enthalten sein. Zum anderen wird eine möglichst *gute Sortierung* verlangt, d.h. die Wortabbilder, die der Anfrage entsprechen, sollten möglichst vor allen anderen Wortabbildern stehen. Weiterhin kann eine *Begrenzung der Länge* der Antwortliste sinnvoll sein.

Konkrete Anwendungen außer der reinen Wortsuche sind unter anderem die Unterstützung beim manuellen Transkribieren[2] oder die automatische Gruppierung [FRG14] oder Indexierung [RM07] von Dokumenten anhand von bestimmten Schlüsselworten.

Seit 1996 ist Word Spotting Forschungsthema im Bereich der Dokumentenanalyse [MHR96]. In seither erschienenen Arbeiten wird zwischen verschiedenen Anfragetypen unterschieden. Beim *Query-by-Example* (QbE) Anfragetyp wird dem Word Spotting System ein Beispielbild des zu suchenden Wortes übergeben. In den Dokumenten wird dann nach Stellen gesucht, die dem Anfragebild optisch ähnlich sind. Ein anderer Anfragetyp ist *Query-by-String* (QbS). Die Anfrage wird in dem Fall als Zeichenkette an das Word Spotting System übergeben. Auch wenn weitere Anfragetypen, wie z.B. *Query-by-Wordclass* [RP09] existieren, so zeigt der letzte Wettbewerb "Keyword Spotting for Handwritten Documents" [PTV15] auf der ICDAR[3] 2015, dass QbE und QbS aktuell die populärsten Anfragetypen sind.

Eine weitere Unterscheidung wird zwischen *trainingsfreiem* und *trainingsbasiertem* Word Spotting vorgenommen. Dabei ist das Training mit annotierten Trainingsdaten gemeint. Auch wenn trainingsbasierte Ansätze bessere Ergebnisse als Trainingsfreie erzielen (vgl. [PTV15]), so lassen sich diese nicht im Fall von fehlenden oder unzureichenden Trainingsdaten anwenden. Gerade bei historischen Dokumenten ist die Erstellung einer Trainingsbasis nicht immer möglich [FRG14]. Die Entwicklung trainingsfreier Word Spotting Systeme ist also gerechtfertigt.

2 Word Spotting - As recommender system to transcription process: http://vc.ee.duth.gr/ws/
3 International Conference on Document Analysis and Recognition

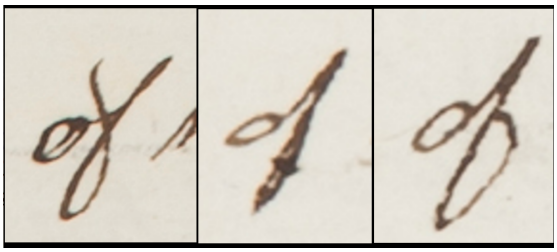

Abbildung 1.0.1: Beispiel für Schwierigkeiten bei der OCR und beim Word Spotting in handschriftlichen Dokumenten. Das Wort "of" wurde an drei verschiedenen Stellen einer Dokumentenseite aus der "Bentham Collection" (siehe Kapitel 5.1) ausgeschnitten.

Für bisher in der Literatur vorgeschlagene Word Spotting Systeme lassen sich oft die folgenden drei Schritte identifizieren:

1. Suche nach Wortkandidaten
2. Repräsentation von Wortkandidaten
3. Vergleich zwischen Anfrage und Wortkandidaten

Im ersten Schritt geht es darum Bildausschnitte zu finden, an denen das Anfragewort vorkommen könnte. Zwei grundsätzliche Vorgehensweisen lassen sich hier finden:
Zum einen gibt es *segmentierungsbasiertes* Word Spotting. Dabei werden die Dokumente in einem Vorverarbeitungsschritt analysiert und in einzelne Wortabbilder segmentiert. Es werden also zunächst alle Bildausschnitte, die genau ein Wort enthalten, als Wortkandidaten angenommen. Zur Anfragezeit werden die segmentierten Wortabbilder mit der Anfrage verglichen. Beispiele für solchen Ansätze sind [RM07] und [RP09].
Zum Anderen gibt es *segmentierungsfreie* Word Spotting Ansätze. Bei diesen wird versucht eine Segmentierung und die damit verbundenen Annahmen über die Struktur der Dokumente zu umgehen. Fehler, die bei der Segmentierung gemacht werden, können in segmentierungsbasierten Systemen nicht wieder gut gemacht werden [RRLF14]. Wenn die Annahmen für die Segmentierungsmethode unpassend sind, kann diese und damit auch das Word Spotting Ergebnis beliebig schlecht werden. Segmentierungsfreie Ansätze können den segmentierungsbasierten Ansätzen also dann überlegen sein, wenn eine Segmentierungsmethode viele Fehler machen würde [GP09].
Aus Gründen der Vollständigkeit sei hier erwähnt, dass es auch segmentierungsbasierte Word Spotting Systeme gibt, die keine Segmentierung auf Wortebene durchführen. In [TT09] wird beispielsweise eine Zeilensegmentierung vorausgesetzt.

Nachdem mögliche Wortkandidaten für die Anfrage gefunden wurden, werden diese im zweiten Schritt geeignet repräsentiert. Dazu werden Merkmale für die jeweiligen Bildausschnitte extrahiert. Neben verschiedenen Merkmalen wurden in der Word Spotting Literatur zwei Arten von Repräsentationen vorgeschlagen. Es gibt *holistische* Repräsentationen, bei diesen wird jedem Wortkandidaten genau ein Merkmalsvektor zugeordnet. Und es gibt *sequentielle* Repräsentationen, bei welchen jedem Bildausschnitt eine Sequenz von Merkmalsvektoren zugewiesen wird (siehe [RP08]). Die Reihenfolge der Sequenz entspricht häufig der Schreibrichtung, so wird die räumliche Struktur der Worte erfasst [RRLF14].

Von der Art der Repräsentation hängt die anschließende Bewertung der Ähnlichkeit ab. Für sequentielle Repräsentationen lassen sich leistungsstarke Methoden der Mustererkennung wie beispielsweise Dynamic-Time-Warping oder Hidden-Markov-Modelle anwenden, mit denen holistische Repräsentationen einige Male übertroffen wurden (z.B. [RM07], [RRLF14]). Im dritten Schritt wird den Wortkandidaten ein *Ähnlichkeitswert* zugewiesen. Dazu werden die jeweiligen Repräsentationen mit der Repräsentation der Anfrage verglichen. Dieser Schritt wird im weiteren Verlauf der Arbeit als *Ähnlichkeitsbewertung* bezeichnet. Die Antwort des Word Spotting Systems ist die nach Ähnlichkeitswert sortierte Liste der Wortkandidaten.

1.2 RAHMEN DER ARBEIT

Word Spotting lässt sich in das Gebiet der Musterkennung einordnen. Ein Teilgebiet der Mustererkennung ist die Nachbildung von Perzeptionsleistungen des Menschen bei speziellen Aufgaben (nach [Nie03] Kapitel 1). Ziel der Mustererkennung ist es mathematisch-technische Verfahren zu entwickeln, um solche speziellen Aufgaben automatisch lösen zu können (nach [Nie03] Kapitel 1, Definition 1.4). Die Verfahren sollten mindestens an die Leistung des Menschen bezüglich der Genauigkeit heranreichen und hinsichtlich der Geschwindigkeit möglichst um ein vielfaches übertreffen. Beim Word Spotting besteht die Perzeptionsleistung des Menschen in der Erkennung aller Vorkommen eines Anfragewortes innerhalb einer Menge von Dokumenten. Wie in der Mustererkennung üblich, ist die Wahl von geeigneten Merkmalen auch für Word Spotting entscheidend und im allgemeinen auf Heuristiken und konkretes Wissen über den Anwendungsbereich angewiesen (siehe auch [Nie03] Kapitel 3). Beim Word Spotting müssen die Merkmale die Unterscheidung zwischen verschiedenen Worten ermöglichen. Deshalb befasst sich diese Arbeit mit dem zuvor beschriebenen Schritt der Repräsentation und der damit verbundenen Merkmalsextraktion. Um den Fokus auf dem Repräsentationsschritt zu halten, wird ein segmentierungsbasiertes Word

Spotting Szenario zur Auswertung vorausgesetzt, ähnlich wie in [ZPG14] und [SF15]. Außerdem wird, um dem Kontext von historischen Dokumenten gerecht zu werden, auf ein trainingsfreies, QbE Word Spotting Szenario gesetzt.
In dieser Arbeit werden zwei in der Literatur vorgeschlagene sequentielle Repräsentationen zusammen mit verschiedenen Histogramm-basierten Merkmalen betrachtet.
Bei der ersten Repräsentation handelt es sich um eine Sequenz von lokalen Merkmalen, welche direkt aus den segmentierten Wortabbildern extrahiert werden. Grundlage für diese Merkmale sind lokale Bilddeskriptoren (D-Sequenz). 2008 wurde eine solche Sequenz für ein segmentierungsbasiertes Word Spotting Szenario vorgeschlagen [RP08].
Bei der zweiten Repräsentation handelt es sich um eine Sequenz von Bag-of-Features (BoF-Sequenz), welche auf lokale Merkmale in einer Gitteranordnung aufsetzt. Diese lokalen Merkmale basieren auf lokalen Bilddeskriptoren. Eine solche Repräsentation wurde zum ersten Mal 2013, für ein segmentierungsfreies Word Spotting Szenario, vorgeschlagen [RRF13], wobei der SIFT-Deskriptor [Low04] als Basis für die lokalen Merkmale diente.
Zur Ähnlichkeitsbewertung wird ein Dynamic-Time-Warping vorgeschlagen. Diese Methode wurde bereits für andere sequentielle Repräsentationen erfolgreich zum Word Spotting eingesetzt (z.B. [RM07]).
Nach dem Wissenstand des Autors zum Abgabezeitpunkt dieser Arbeit ist dies die erste Arbeit, in der eine BoF-Sequenz in Kombination mit Dynamic-Time-Warping zum Word Spotting untersucht wird. Die Untersuchung dieser Kombination ist daher ein Teilaspekt dieser Arbeit.
Der direkte Vergleich zwischen der moderneren Bag-of-Features Sequenz (BoF-Sequenz) und einer Sequenz von lokalen Bilddeskriptoren (D-Sequenz) ist ebenfalls ein Beitrag der vorliegenden Arbeit. Die **Leitfrage** dieser Arbeit ist, welche der beiden sequentiellen Repräsentationen besser zum Word Spotting in historischen Dokumenten geeignet ist.

1.3 AUFBAU DER ARBEIT

Der weitere Teil dieser Bachelorarbeit ist wie folgt strukturiert: In Kapitel 2 werden die Grundlagen, die für das Verständnis der Arbeit essentiell sind, erklärt. In Kapitel 3 wird auf Arbeiten, welche diese Arbeit beeinflusst haben, eingegangen. In Kapitel 4 wird die konkrete Word Spotting Methodik dieser Arbeit vorgestellt. Eine Auswertung dieser Methodik ist in Kapitel 5 zu finden. Kapitel 6 umfasst eine Zusammenfassung der wichtigsten Ergebnisse sowie ein Fazit dieser Arbeit.

2 GRUNDLAGEN

Zur Verarbeitung historischer Dokumente mit einem Computer müssen diese zunächst digitalisiert werden. Dazu existieren eine Reihe von Techniken und Standards (siehe z.B. [Con06]). Anschließend liegen die Dokumente als digitale Bilder vor. Ein *digitales Bild* lässt sich durch eine diskrete *Bildmatrix* darstellen. Die Einträge der Bildmatrix werden *Bildpixel* und die Größe $X \times Y$ der Matrix wird *Bildauflösung* genannt. Man kann zwischen verschiedenen *Bildtypen* unterscheiden, je nach Bildtyp sind die Bildpixel Elemente eines anderen Wertebereichs mit verschiedenen Interpretationen (siehe auch [Pri15] Kapitel 4.1). Bei digitalisierten Dokumentenabbildern handelt es sich üblicherweise um *Farb- oder Graustufenbilder*. Es sei angemerkt, dass es auch Ausnahmen geben kann (z.B. [RFM+15]). In dieser Arbeit wird davon ausgegangen, dass die Dokumentenabbilder zunächst als Farb- oder Graustufenbild vorliegen.
In der Wort Spotting Literatur dienen häufig andere Bildtypen als Grundlage für die Repräsentation der Wortkandidaten. Frühe Ansätze setzten beispielsweise auf *Binärbilder* (z.B. [MHR96]), modernere Ansätze setzen auf *Gradientenbilder* (z.B. [RP09], [RATL11]). In dieser Arbeit sind beide genannten Bildtypen für den Schritt der Repräsentation relevant, daher werden diese zunächst in 2.1 und 2.2 erklärt. Anschließend wird in Unterkapitel 2.3 auf weitere Grundlagen eingegangen, welche für die beiden sequentiellen Repräsentationen dieser Arbeit und das Verständnis verwandter Arbeiten relevant sind. In Abschnitt 2.4 werden die Grundlagen für die Ähnlichkeitsbewertung dieser Arbeit erläutert. Zum Abschluss des Kapitels erfolgt eine Übersicht über die angesprochenen Themen sowie Anknüpfungspunkte an weitere Literatur.

2.1 BINARISIERUNG

Den Schritt der Erzeugung von Binärbildern aus Farb- bzw. Graustufenbildern nennt man Binarisierung. Ziel der Binarisierung von Dokumentenabbildern ist die Unterteilung in Schriftpixel (Vordergrund) und Hintergrundpixel. Intuitiv beschreiben Binärbilder Schrift nicht schlechter als Graustufenbilder, für die Weiterverarbeitung reduziert sich jedoch der Speicher- und Rechenaufwand (siehe auch [Nie03] Kapitel 2.2.1).
Ausgehend von Graustufenbildern werden häufig Schwellenwertverfahren eingesetzt,

um eine Binarisierung durchzuführen. Dabei wird für jeden Bildpixel ein Schwellwert bestimmt. Je nachdem ob der Grauwert des Bildpixels über oder unter dem Schwellwert liegt, wird die Einteilung in Vordergrund bzw. Hintergrund vorgenommen. Man unterscheidet zwischen globalen und lokalen Schwellwertverfahren. Bei globalen Verfahren wird für jeden Bildpixel, aufgrund einer globalen Analyse des Bildes, der gleiche Schwellwert angenommen. Bei lokalen Schwellwertverfahren werden unterschiedliche Schwellwerte, durch Analyse einzelner Bildausschnitte, verwendet. Bei der Binarisierung von historischen Dokumentenabbildern sind lokale Schwellwertverfahren gegenüber globalen im Vorteil (nach [SPoo]). Alterungserscheinungen in Teilen eines Dokuments können den Schwellwert verfälschen, bei einem globalen Verfahren würde so die Binarisierung des ganzen Dokumentes scheitern. Lokale Verfahren erzeugen nur lokale Fehler.

Niblack-Schwellwert

Für Dokumentenabbilder eignet sich die von Niblack vorgeschlagene Methode zur Bestimmung eines lokalen Schwellwertes. Der Niblack-Schwellwert T_N für einen Bildausschnitt wird wie folgt berechnet: $T_N = \bar{G} + \alpha * S_G$ (siehe [Nib86]). Wobei $\bar{G}$ der Mittelwert und S_G die Standardabweichung der Grauwerte des betrachteten Bildausschnittes sind. $\alpha <= 0$ ist ein frei wählbarer Parameter. Betrachtet man die Häufigkeitsverteilung der Grauwerte für einen Bildausschnitt aus einem Dokumentenabbild, so enthält dieser normalerweise deutlich mehr Hintergrundpixel. Der Mittelwert tendiert also in Richtung des Hintergrundes, die Standardabweichung beschreibt dann in etwa die Abweichung vom Hintergrund. Durch den Parameter α lässt sich der Schwellwert so anpassen, dass die Grenze zwischen Schrift und Hintergrund passend bestimmt werden kann. Abbildung 2.1.1 soll dies anhand eines Beispiels verdeutlichen.

2.2 BILDGRADIENTEN

Bildgradienten lassen sich zur Kantendetektion einsetzen (siehe [Jä12] Kapitel 12.4). Bildgradienten sind daher als Grundlage zur Charakterisierung von Schrift besonders geeignet. Dies wird intuitiv, spätestens am Ende dieses Unterkapitels, deutlich.
Betrachtet man die Bildgradienten erster Ordnung, also die erste Ableitung, $\nabla I(x, y)$ eines kontinuierlichen Bildsignal $I(x, y)$, so stellt man fest, dass ihre Extremstellen

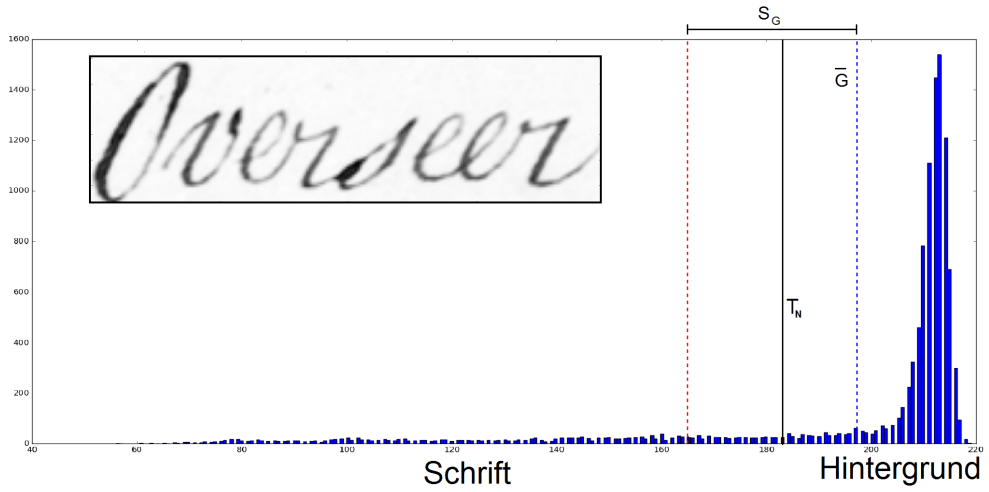

Abbildung 2.1.1: Beispiel für den Niblack-Schwellwert für ein Wortabbild des Wortes "overseen".

an den Stellen zu finden sind, an denen Kanten im Bild vorhanden sind (siehe [Jä12] Kapitel 12.1 und 12.2).

$$\nabla I(x,y) = \begin{pmatrix} \frac{\partial I(x,y)}{\partial x} \\ \frac{\partial I(x,y)}{\partial y} \end{pmatrix} \tag{2.2.1}$$

Im diskreten Bildsignals $I_d(x,y)$ kann eine Approximation der Bildgradienten erster Ordnung durch eine diskrete Faltung mit einem *Gradienten-Filter* in x- bzw. y-Richtung erreicht werden (siehe [Pri15] Kapitel 6.4.1, Seite 130). Man erhält anschließend ein diskretes Gradientenbild G_d, bei dem die Bildpixel 2-dimensionale Vektoren sind. Die Elemente dieser Vektoren sind die durch die beiden Faltungen erzeugten Werte.

$$G_d(x,y) = \begin{pmatrix} g_x \\ g_y \end{pmatrix} \tag{2.2.2}$$

Die einzelnen Bildgradienten lassen sich durch ihre Magnitude (Stärke) M und Orientierung θ charakterisieren.

$$m(x,y) = \sqrt{g_x^2 + g_y^2}, \qquad \theta(x,y) = \measuredangle(g_x, g_y) \qquad \text{nach [RP08]} \tag{2.2.3}$$

Die Magnitude eines solchen Vektors entspricht also seiner Länge. Die Orientierung lässt sich als Winkel zu einer Bezugsachse im kartesischen Koordinatensystem angeben.

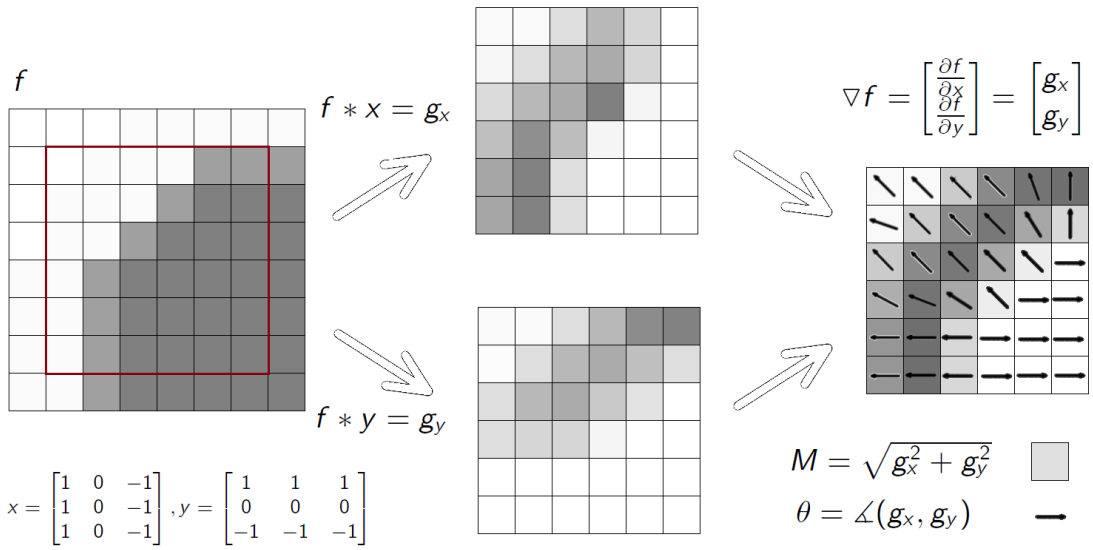

Abbildung 2.2.1: Schema der Berechnung von Bildgradienten mit Prewitt-Filter, anhand eines Beispiels. Links: Graustufenbild f. Mitte: Die jeweiligen Bilder, welche mit den aufgeführten Filter-Masken erzeugt wurden. Rechts: Das Gradientenbild, wobei die Magnitude als Grauwerte und Orientierung als Pfeile dargestellt sind. Die Filterbilder und das Gradientenbild sind nur für den im Graustufenbild rot umrahmten Bereich dargestellt.

Dazu lässt sich die Umkehrfunktion des Tangenz benutzen (siehe [RP08]). Je stärker eine Kante im Bild, desto größer ist die Magnitude. Die Orientierung des Gradienten entspricht der Orientierung orthogonal zur Kante. Abbildung 2.2.1 verdeutlicht das Berechnungsschema für Bildgradienten in digitalen Bildern und zeigt beispielhaft die Magnituden und Orientierung, die bei der Berechnung entstehen. Für die Gradientenberechnung wurden in der Literatur einige Gradienten-Filter vorgeschlagen, welche verschiedene Eigenschaften besitzen (siehe [Pri15] Kapitel 6.4.1). Zwei bekannte Gradienten-Filter zur Kantendetektion sind der *Prewitt-Filter* und der *Sobel-Operator*. Diese besitzen zusätzliche Glättungseigenschaften durch die Einbeziehung von benachbarten Pixeln. Das macht sie robust gegen leichtes Bildrauschen (siehe [Pri15] Kapitel

6.4.1). Bei der Berechnung der Bildgradienten wird hier deshalb der Sobel-Operator eingesetzt:

$$S_X = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}, S_Y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \text{ (nach [DH73])} \tag{2.2.4}$$

In Abbildung 2.2.2 sind die Bildgradienten eines Wortabbildes visualisiert, um die Eignung von Bildgradienten im Bereich der Analyse von Dokumentenabbildern zu demonstrieren. Die Kantenverläufe der Schrift sind deutlich zu erkennen und klar vom Hintergrund abgehoben.

2.3 EXTRAKTION LOKALER MERKMALE

Wie bereits in der Einleitung 1.2 erwähnt ist das Ziel der Merkmalsextraktion beim Word Spotting Merkmale bzw. Repräsentationen für Wortkandidaten zu finden, für gleiche Wörter ähnlicher bewertet werden als ungleiche Wörter. Wie auch bei anderen Mustererkennungsaufgaben werden dabei zwei Ziele verfolgt, zum Einen die Konzentration auf relevante Informationen und zum Anderen die Reduktion der Datenmenge (nach [Nie03] Kapitel 3.1).
Der weitere Teil dieses Unterkapitels gliedert sich wie folgt. Es wird auf lokale Bilddeskriptoren eingegangen, da diese für beide sequentiellen Repräsentationen dieser Arbeit relevant sind. In der BoF-Sequenz dienen sie als Grundlage zur Bildung *lokaler Merkmale*. In der D-Sequenz werden sie direkt als *Merkmalsvektoren* lokaler Merkmale interpretiert. An dieser Stelle sei ein Verweis auf [Nie03] Kapitel 3 gegeben, dort werden die vorigen Begrifflichkeiten detailliert und allgemein dargestellt.
Weiterhin werden hier zwei Strategien zur Anordnung solcher lokalen Bilddeskriptoren angesprochen. Anschließend wird das Gradientenhistogramm erklärt, welches eine wichtige Grundlage für den SIFT-Deskriptor ist. Dieser dient in der vorliegenden Arbeit als Grundlage für eine Variante der BoF-Sequenz. Nachdem der SIFT-Deskriptor erklärt wurde, wird auf das Bag-of-Features Konzept und das damit verbundene Clustering eingegangen.
Zuvor sei noch das Konzept der Quantisierung angeführt, welches hier grundlegend für das Erreichen des Ziels der Merkmalsextration ist.

Abbildung 2.2.2: Gradientenbild mit Sobel-Operator. Oben links: Graustufenbild. Unten links: Bild der Magnituden, Darstellung im Grautwertbereich. Je heller desto größer die Magnitude. Oben rechts: Bild der Orientierungen im Bogenmaß. Darstellung im HSV-Farbraum (siehe auch [Pri15] Kapitel 3.2.4). Jede Farbe steht für eine Orientierung, ähnliche Farben entsprechen ähnlichen Orientierungen. Unten rechts: Kombination der Magnituden und Orientierungen. Darstellung im HSV-Farbraum.

Quantisierung

Ziel einer Quantisierung ist es Daten aus einem Eingabedatum auf eine (kleinere) endliche Menge typischer Repräsentanten abzubilden, ohne dass dabei für die weitere Verarbeitung relevante Information verloren geht (nach [Fin14] Seite 51). Durch die Quantisierung lässt sich neben der Reduktion der Daten auch eine gewisse Abstraktion herstellen, welche nötig ist, um mit Schriftvariationen umgehen zu können.

2.3.1 *Lokale Bilddeskriptoren*

Lokale Bilddeskriptoren dienen hier der Beschreibung kleinerer Bildausschnitte der Wortkandidaten. Die konkreten Ausprägungen der Merkmale hängen also jeweils nur von einem lokalen Bereich ab und werden nicht durch Variationen an anderen Stellen des Wortkandidaten beeinflusst. Für die Auswahl solcher Bildausschnitte sind zwei Strategien in dieser Arbeit relevant.
Die Eine wird als *dichtes Gitter* bezeichnet und ist vor allem für Bag-of-Features Ansätze zum Word Spotting relevant. Dabei werden Bilddeskriptoren mit festgelegter Größe in einem Gitter über die Dokumentenabbilder bzw. Wortkandidaten angeordnet (siehe [RATL11], [RRLF14], [SF15]). Die Bereiche der Deskriptoren überschneiden sich dabei.

Die andere Strategie ist die Benutzung eines *gleitenden Fensters*. Dabei wird ein Fenster mit einer bestimmten Breite in regelmäßigen Abständen in Schreibrichtung über einen Wortkandidaten bewegt. Auch hier überlappen sich die Bildausschnitte. An jeder Fensterposition wird der Bildausschnitt analysiert und ein Deskriptor berechnet. Durch die Bewegung des Fensters in Schreibrichtung entspricht die Folge der Deskriptoren annähernd der zeitlichen Abfolge beim ursprünglichen Schreibprozess (vgl. [RP08]).
Der Vollständigkeit halber sei hier erwähnt, dass es in der Literatur auch andere Strategien für die Auswahl solcher lokalen Bereiche in der Bilderkennung und zum Word Spotting gibt. So findet man beispielsweise Ansätze, bei denen vorher bestimmte Punkte ausgewählt werden, welche sich besonders für die Charakterisierung der Wortkandidaten eignen sollen (z.B. [AFV13]).
Einer der erfolgreichsten Bilddeskriptoren der Bilderkennung ist der SIFT-Keypoint Deskriptor, der ursprünglich zur Objekterkennung in Bildern vorgeschlagen wurde. Dabei wird er zur Beschreibung vorher ermittelter Interessenspunkte (Keypoints) mit einer bestimmten Orientierung und Skalierung eingesetzt (siehe [Low04]).

Gradientenhistogramm

Grundlage für den SIFT-Keypoint Deskriptor sind Gradientenhistogramme. Zunächst wird in diesem Abschnitt auf den Begriff des Histogramms eingegangen. Anschließend werden die Details zum Gradientenhistogramm erklärt.
In der Literatur lassen sich diverse Definitionen zu verschiedensten Arten von Histogrammen finden, die Definitionen reichen dabei meist nur so weit, wie es der Kontext erfordert. So werden beispielsweise in ([Pri15] Kapitel 4.3) diverse Definitionen für Histogramme gegeben. Für eine allgemeinere Definition sei auf ([Nie03] Kapitel 4.2 S.327 ff.) verwiesen. Die wesentlichen Punkte dieser Definitionen sind jeweils das Zählen der Werte innerhalb einer Stichprobe. Weiterhin wird der mögliche Wertebereich dieser Werte häufig in mehrere gleich große Klassen eingeteilt (nach [Nie03] Kapitel 4.2 S.327 ff.). Ein Histogramm stellt dann die Anzahlen der Werte jeder Klasse innerhalb einer Stichprobe dar. So lässt sich mit einem Histogramm beispielsweise die Häufigkeitsverteilung der Grauwerte eines Bildes darstellen. Ein solches *Grauwerthistogramm* ist in Abbildung 2.1.1 zu sehen. Die Klassen entsprechen dabei jeweils einem diskreten Grauwert.
In einem *Gradientenhistogramm* werden die Bildgradienten eines Bildausschnittes zusammengefasst. Die Klasseneinteilung erfolgt dabei über die Orientierung der Bildgradienten. Man unterteilt den Orientierungsbereich in eine bestimmte Anzahl T gleichgroßer Klassen. Um die Information der Magnitude nicht zu verlieren, wird in den Klassen nicht die Anzahl der Bildgradienten erfasst, sondern die summierten

Werte der Magnituden der Bildgradienten dieser Klasse. Eine Klasse entspricht dabei jeweils einer *Hauptorientierung*.
Für die Berechnung eines Gradientenhistogramms werden die Bildgradienten zunächst auf diese Hauptorientierungen quantisiert (vgl. [RP08]). Für die Quantisierung der Bildgradienten auf die Hauptorientierungen gibt es verschiedene Verfahren. Für diese Arbeit ist die in [RP08] vorgeschlagene *lineare Interpolation* der Bildgradienten auf die jeweils benachbarten Hauptorientierungen relevant. Für jeden Bildgradienten wird die Magnitude linear auf die benachbarten Hauptorientierungen aufgeteilt. Dadurch wird das Aliasrauschen bei der Quantisierung abgeschwächt (vgl. [RP08]). Abbildung 2.3.1 zeigt schematisch wie das Gradientenhistogramm zu den vorherigen Bildausschnitt f (siehe Abbildung 2.2.1) aussehen könnte. Außerdem wird die zuvor genannte lineare Interpolation skizziert. Man sieht, dass das Gradientenhistogramm den Kantenverlauf von f abstrahiert darstellt.

SIFT-Keypoint Deskriptor

Im Bereich Word Spotting wird der SIFT-Keypoint Deskriptor [Low04] häufig in Kombination mit dem später erläuterten Bag-of-Features Ansatz eingesetzt. Dabei werden mehrere dieser Deskriptoren in einem dichten Gitter angeordnet. Der Aspekt der Keypoints ist hier also weniger relevant. Im Folgenden wird der SIFT-Keypoint Deskriptor daher nur noch als SIFT-Deskriptor bezeichnet. Da die Schrift in den später untersuchten Dokumenten nicht stark gedreht ist, ist es sinnvoll für alle Deskriptoren die gleiche Orientierung vorauszusetzen (siehe auch [RRF13]). Jeder SIFT-Deskriptor im dichten Gitter setzt sich wie folgt zusammen:
Für den betrachteten Bildausschnitt wird das Gradientenbild betrachtet. Die Magnituden der einzelnen Bildgradienten werden mit einem Gauß-Fenster gewichtet. Dadurch fallen die Bildgradienten, die weiter vom Mittelpunkt des Ausschnittes entfernt sind, weniger ins Gewicht. Weiterhin wird der Bildbereich in eine feste Anzahl von $c \times c$ gleichgroße Zellen unterteilt, dadurch erfasst der Deskriptor räumliche Informationen um den Mittelpunkt. Für jede dieser Zellen wird ein Gradientenhistogramm berechnet. Die Konkatenation der Gradientenhistogramme stellt den Deskriptor dar. Man erhält also für jede Gitterposition einen Vektor mit $c \times c \times T$ Dimensionen. In der Praxis wird häufig eine Zellenaufteilung von 4×4 und $T = 8$ Hauptrichtungen verwendet [Low04].
Weitere für diese Arbeit relevante lokale Bilddeskriptoren werden im Kapitel 4.3 näher erläutert.

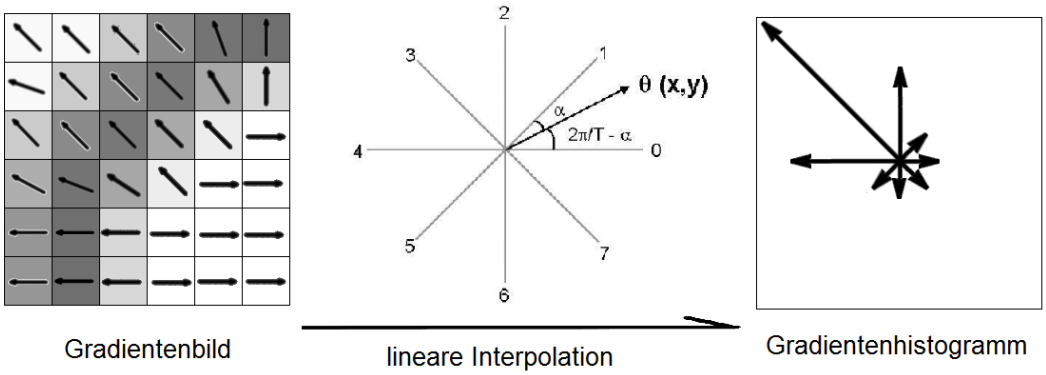

Abbildung 2.3.1: Beispiel für die Bildung eines Gradientenhistogramm mit T = 8 Hauptorientierungen und linearer Interpolation. Lange Pfeile im Gradientenhistogramm entsprechen hohen Werten. Die mittlere Grafik stammt aus [RP08].

2.3.2 *Bag-of-Features*

Bag-of-Features sind Histogramme über die Häufigkeiten charakteristischer Merkmale, die durch ein sogenanntes Codebuch vorgegeben sind. Ursprünglich stammen Bag-of-Features aus der Analyse maschinenlesbarer Dokumente, dort werden diese Bag-of-Words genannt. Die Merkmale des Codebuches sind in dem Fall Worte bzw. Wortstämme, die charakteristisch für bestimmte Kategorien von Dokumenten sind. Die Bag-of-Words eines Dokuments enthält dann die Häufigkeiten der Worte bzw. Wortstämme aus dem Codebuch innerhalb des Dokuments. Mit den Bag-of-Words ist eine einfache Kategorisierung von Dokumenten möglich [OD11].
Das gleiche Konzept hat sich auch im Bereich der Mustererkennung in Bildern durchgesetzt. Aufgrund der Ähnlichkeit zu den Bag-of-Words werden die Bag-of-Feautres hier auch häufig als Bag-of-Visual-Words bezeichnet. Das Codebuch wird dann *visuelles Vokabular* und die Einträge des Codebuches *visuelle Wörter* genannt. Das visuelle Vokabular für Bag-of-Visual-Words wird in der Regel durch ein Clustering von aus Trainingsbildern extrahierten lokalen Bilddeskriptoren erzeugt [OD11].
Eine ausführliche Abhandlung zum Bag-of-Features Prinzip für Bildklassifikation und Bildretrieval ist in [OD11] zu finden.
Im Folgenden werden die Aspekte erklärt, die für das Verständnis der in Kapitel 3.2 aufgeführten Arbeiten und für diese Arbeit wichtig sind.
Eine Methode zur Extraktion von Merkmalen für eine Bag-of-Features Repräsentation zum Word Spotting wurde bereits in Abschnitt 2.3.1 angesprochen. Typischerweise

werden lokale Bilddeskriptoren in einem dichten Gitter berechnet und diese anschließend auf ein visuelles Vokabular quantisiert. Die lokalen Bilddeskriptoren, die für das Erstellen des visuellen Vokabulars benötigt werden, werden in einem Vorverarbeitungsschritt auf einer geeignet großen Trainingsmenge von Dokument- bzw. Wortabbildern ebenfalls in einem dichten Gitter berechnet (siehe z.B. [RATL11], [RRLF14], [ARTL15]).

Clustering

Durch Clustering lassen sich typische Repräsentanten für Häufungsgebiete ähnlicher Datenpunkte in einer größeren Datenmenge finden, die sich z.B. für eine spätere Quantisierung eignen (siehe [Fin14] Seite 51 ff.).

Lloyd-Arlgorithmus

In dieser Arbeit wird eine Variante des Lloyd-Algorithmus zum Clustering eingesetzt. Der Lloyd-Algorithmus erzeugt durch iterative Verbesserung, ausgehend von einer initialen Clusterkonfiguration, eine lokal optimale Clusterkonfiguration. Eingabe für den Algorithmus sind eine geeignet große Stichprobe von Daten, die gewünschte Anzahl x an Codebucheinträgen und eine Abbruchschranke. Der Algorithmus lässt sich in die folgenden 4 Schritte unterteilen:

1. **Initialisierung**: Es werden zufällig x Datenpunkte aus der Stichprobe als initiale Cluster Repräsentanten bzw. Codebucheinträge ausgewählt.

2. **Verbesserung**: Alle Datenpunkte werden zu ihrem nächsten Nachbarn aus dem Codebuch zugeordnet, wodurch eine neue Verteilung der Datenpunkte in die verschiedenen Cluster entsteht.

3. **Aktualisierung**: Die neuen Mittelpunkte jedes Clusters werden berechnet und als neue Repräsentanten bzw. Codebucheinträge festgelegt.

4. **Abbruchkriterium**: Wenn die Abbruchschranke überschritten wird, bricht der Algorithmus ab und gibt das aktuelle Codebuch zurück. Wenn nicht, wird mit Schritt 2 fortgefahren.

Für eine ausführlichere Einführung zum Thema Quantisierung und Clustering sei auf weitere Literatur verwiesen, z.B. Kapitel 4.3 aus [Fin14].

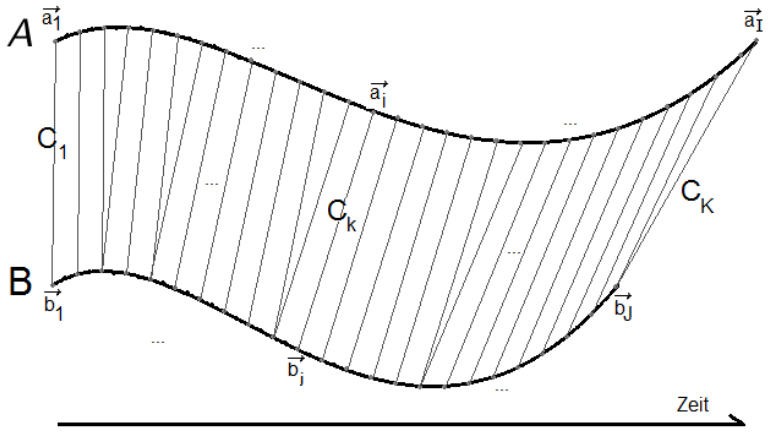

Abbildung 2.4.1: Beispiel für einer Warping Funktion für zwei Sequenzen eindimensionaler Merkmalsvektoren. Die Verbindungslinien visualisieren die Warping Funktion.

2.4 DYNAMIC-TIME-WARPING

Dynamic-Time-Warping (DTW) ist eine Methode zur Berechnung einer Distanz zwischen zwei Sequenzen von Merkmalsvektoren. Ursprünglich wurde DTW im Bereich der Spracherkennung für den Vergleich zwischen zeitabhängig generierten Merkmalssequenzen von Sprachmustern entwickelt [SC78]. In solchen Sprachmustern besteht häufig das Problem, dass es nicht lineare Zeitunterschiede zwischen gleichen Sprachmustern gibt, z.B. wegen unterschiedlicher Sprechgeschwindigkeiten. Mit dem DTW wird versucht solche Zeitvariationen auszugleichen. Seitdem wurde DTW auch erfolgreich in anderen Bereichen der Mustererkennung eingesetzt, unter anderem zum Word Spotting (siehe Kapitel 3). Hier ist DTW in der Lage Variationen in Schreibrichtung der Worte auszugleichen (vgl. [RM07]).
Grundlage für das DTW ist das Model der optimalen Warping Funktion [SC78]. Eine Warping Funktion $F = C_1, ..., C_k, ..., C_K$ ist eine Zuordnung der Merkmalsvektoren zweier Sequenzen A und B von Merkmalsvektoren, $A = \vec{a_1}, ..., \vec{a_i}, ..., \vec{a_I}$ und $B = \vec{b_1}, ..., \vec{b_j}, ..., \vec{b_J}$, zueinander (vgl. [SC78]). Abbildung 2.4.1 zeigt ein Beispiel für eine mögliche Warping Funktion zwischen zwei Sequenzen eindimensionaler Merkmalsvektoren.
Unter allen denkbaren Warping Funktionen ist nach [SC78] diejenige Warping Funktionen optimal, für welche die gewichtete Summe der Distanzen aller Zuordnungspaare

minimal ist. Die Warping Funktion in Abbildung 2.4.1 sieht nach diesem Kriterium bereits nach einer guten Warping Funktion aus. Es kann angenommen werden, dass diese minimale Summe der verbleibenden Distanz zwischen den Sequenzen entspricht, wenn es keine Zeitvariation zwischen diesen gibt [SC78]. Sie wird deshalb als zeitnormalisierte Distanz bezeichnet und ist definiert durch:

$$D(A,B) = \min_{F}\left(\frac{\sum_{k=1}^{K} d(C_k)\cdot w(k)}{\sum_{k=1}^{K} \cdot w(k)}\right) \quad \text{nach [SC78]} \tag{2.4.1}$$

$w(k)$ ist als einstellbarer Gewichtungsfaktor entworfen worden, um es zu ermöglichen verschiedenen Charakteristiken der Daten erfassen zu können [SC78]. Und $d(C_k)$ ist ein beliebiges Distanzmaß zur Berechnung der Distanz zwischen den beiden Merkmalsvektoren von C_k.
Sowohl für den praktikablen Einsatz, als auch für die Berechnung sind fünf Bedingungen an die Warping Funktion zu stellen [SC78].

1. **Monotonie**: Es wird nicht in der Zeit bzw. den Sequenzen zurück gegangen.
2. **Kontinuierlichkeit**: Es wird kein Merkmalsvektor ausgelassen.
3. **Begrenzung**: Es werden jeweils die ersten und die letzten beiden Merkmale einander zugeordnet.
4. **Anpassungsfenster**: Die Menge der infrage kommenden Warping Funktionen kann durch ein Fenster beschränkt werden.
5. **Steilheitseinschränkung**: Der Warping Pfad kann in seiner lokalen Steilheit eingeschränkt werden.

Mit Hilfe dieser Bedingungen lassen sich DTW-Algorithmen formulieren, wobei die Steilheitseinschränkung maßgeblich für den konkreten Algorithmus ist [SC78]. In Kapitel 4.4 wird auf die Details des DTW-Algorithmus dieser Arbeit eingegangen. Abbildung 2.4.2 skizziert das Konzept der Warping Funktion zusammen mit einigen der genannten Bedingungen. Das dargestellte Anpassungsfenster, welches in der Abbildung zu sehen ist, wird in der Literatur auch als "Sakoe-Chiba Band" bezeichnet (siehe [RM03]). Die zeitnormalisierte Distanz dient in dieser Arbeit als Ähnlichkeitswert zwischen Anfrage und Wortkandidaten.

Distanzmaße

Vergleicht man Merkmalsvektoren miteinander, so ist es wichtig ein passendes Distanzmaß zu verwenden. Dass der Einsatz eines unpassenden Distanzmaßes beim Word

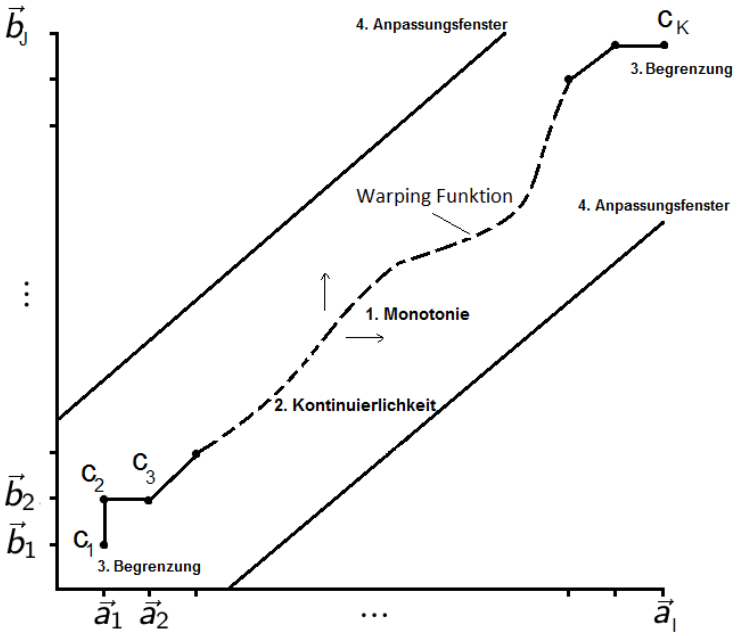

Abbildung 2.4.2: Konzept der Warping Funktion und Eigenschaften 1-4. (Abbildung nach [SC78])

Spotting mit einer Bag-of-Features basierten Repräsentation zu schlechten Resultaten führt, wurde beispielsweise in [SF15] demonstriert. Das Distanzmaß muss also zu den Eigenschaften der Merkmale bzw. Merkmalsvektoren passen. Für den Vergleich von hochdimensionalen Vektoren im ersten Quadranten mit vielen Nulleinträgen wird häufig die *Kosinus-Distanz* verwendet [Sino1]. Daher wird die Kosinus-Distanz häufig für den Vergleich von Bag-of-Features basierten Repräsentationen genutzt (z.B. [RATL11]). In [SF15] wurde demonstriert, dass die *BrayCurtis-Distanz* unter bestimmten Bedingungen bessere Resultate liefert als die Kosinus-Distanz.

$$BC(\vec{a}, \vec{b}) = \frac{\sum_i | a_i - b_i |}{\sum_i a_i + b_i} \tag{2.4.2}$$

Bei der BC-Distanz ist zu beachten, dass diese der *L1-Distanz* entspricht, wenn die Summen der Komponenten der jeweilige Merkmalsvektoren der gleichen Konstante c entsprechen. Dann gilt: $BC(\vec{a}, \vec{b}) = \frac{1}{2c} L1(\vec{a}, \vec{b})$.
Für den Vergleich zweier Histogramme erscheint die L1-Distanz intuitiv passend, da sie die Differenzen, also die Unterschiede, der einzelnen Einträge im Histogramm

direkt erfasst.
Die vierte Distanz, welche in dieser Arbeit betrachtet wird, ist die *L2-Distanz*, auch euklidische Distanz, da diese ebenfalls ein bekanntes Distanzmaß zwischen Vektoren ist.

ZWISCHENÜBERSICHT

Zu Beginn des Kapitels wurde der Grundbegriff des *digitalen Bildes* geklärt, da dies die Darstellungsgrundlage für die Dokumentenabbilder ist. Anschließend wurden die Bildtypen *Binärbild* und *Gradientenbild* erläutert, da die weitere Merkmalsextraktion hier auf diese beiden Bildtypen aufbaut.
Anschließend wurde auf die Extraktion lokaler Merkmale eingegangen und damit verbundene Konzepte wie die *Quantisierung* und *lokale Bilddeskriptoren* erläutert. Im Zusammenhang der lokalen Bilddeskriptoren wurden *Histogramme* allgemein und das *Gradientehistogramm* besprochen. Als einer der bekanntesten lokalen Bilddeskriptoren wurde der *SIFT-Deskriptor* vorgestellt. Als Ergänzung sei hier angemerkt, dass nach dem SIFT-Deskriptor in der Literatur weitere Deskriptoren folgten, die das Konzept des Gradientenhistogramms in Kombination mit räumlichen Beschreibung durch eine Gitterstruktur aufgriffen. Ein bekannter Vertreter ist der HOG-Deskriptor, der 2005 zur Menschenerkennung in Bildern vorgeschlagen wurde [DT05]. Derivate des HOG-Deskriptors finden sich auch in Arbeiten zum Thema Word Spotting, beispielsweise in [TT09] oder [AFV13].
Anschließend wurde ein Einstieg zum *Bag-of-Features* Konzept und dem damit verbundenen *Clustering* gegeben.
Zuletzt wurden Grundlagen zu Techniken der Ähnlichkeitsbewertung vorgestellt. Zum einen das *Dynamic-Time-Warping*, dass in dieser Arbeit benutzt wird. Und zum anderen einige *Distanzmaße*, die, neben dem Einsatz in der *zeitnormalisierte Distanz*, auch für den direkten Vergleich zwischen holistischen Wortrepräsentationen genutzt werden können (z.B. [RATL11]). An dieser Stelle sei erwähnt, dass in der Literatur weitere Methoden zur Ähnlichkeitsbewertung beim Word Spotting zu finden sind. Eine weitere Methode für den Vergleich sequentieller Wortrepräsentationen sind beispielsweise Hidden Markov Modelle (z.B. [RP09], [RRF13]), auch Methoden, die auf neuronalen Netzen basieren sind, zu finden (z.B. [FFMB12]).

3 VERWANDTE ARBEITEN

Nachdem die wichtigsten fachlichen Grundlagen erklärt wurden, werden in diesem Kapitel einige Arbeiten, welche im engeren Zusammenhang mit der vorliegenden Arbeit stehen, vorgestellt und diskutiert. Im Vergleich zur Bag-of-Features Sequenz (BoF-Sequenz) handelt es sich bei der Deskriptor Sequenz (D-Sequenz) um einen älteren Ansatz zum Word Spotting (vgl. Kapitel 1.2). Daher werden die Arbeiten zum Word Spotting hier in chronologischer Reihenfolge angeführt.
Zunächst wird allerdings die Arbeit vorgestellt, welche das DTW zum Word Spotting motiviert. Anschließend werden in Abschnitt 3.1 zwei Arbeiten vorgestellt, welche die Grundlage für die D-Sequenz und die Merkmale innerhalb dieser Repräsentation darstellen. In Abschnitt 3.2 werden modernere Bag-of-Features basierte Ansätze zum Word Spotting besprochen. Einige Arbeiten sind dabei grundlegend für die BoF-Sequenz dieser Arbeit. Andere werden für den späteren Vergleich zum State-of-the-Art (siehe Kapitel 5.4) erläutert. Am Ende des Kapitels steht eine Zusammenfassung der wichtigsten Aspekte für diese Arbeit. Zudem wird eine detaillierte Übersicht zu den Untersuchungsaspekten gegeben, welche sich aus den verwandten Arbeiten für die Beantwortung der Leitfrage (siehe Kapitel 1.2) der vorliegenden Arbeit ergeben.

2003 haben Rath und Manmatha in "Word Image Matching Using Dynamic Time Warping" [RM03] einen DTW-Algorithmus zur Ähnlichkeitsbewertung zwischen sequentiellen Repräsentationen segmentierter Wortabbilder vorgeschlagen. Bei der vorgeschlagenen Wortrepräsentation handelt es sich um eine Sequenz von spaltenweise extrahierten Merkmalsvektoren. Die einzelnen Merkmale werden Projection Profile, Upper/Lower Word Profile und Background/Ink Transitions genannt. Der Erfolg dieser Merkmale ist direkt an eine gute Vorverarbeitung geknüpft. So werden vor der Extraktion dieser Merkmale Schriftneigung und Grundlinienversatz in den Wortabbildern korrigiert, um die Schriftvariation in vertikaler Richtung zu reduzieren [RM03]. Der DTW-Algorithmus ist in der Lage Variationen in horizontaler Richtung bei der Ähnlichkeitsbewertung auszugleichen [RM03].

3.1 HISTOGRAMM-BASIERTE MERKMALE

Seit dem Erfolg des SIFT-Deskriptors sind in der Literatur zur Bilderkennung weitere Deskriptoren vorgeschlagen worden, welche auf Gradientenhistogrammen basieren. Auch in der Literatur zum Word Spotting haben solche Deskriptoren schnell Einzug gehalten.

3.1.1 *Local Gradient Histogram Features*

2008 haben Rodriguez und Perronnin in ihrer Arbeit "Local gradient histogram features for word spotting in unconstrained handwritten documents" [RP08] eine sequentielle Wortrepräsentation für ein segmentierungsbasiertes Word Spotting Szenario vorgeschlagen. Die Arbeit war Bestandteil eines Forschungsprojektes, bei dem es darum ging, Schlüsselwörter in einer Menge von gescannten Briefen zu erkennen, beispielsweise für eine automatische Postverteilung innerhalb eines Unternehmens. Es handelt es sich also um ein Mehrschreiber Szenario, bei dem die Variabilität gleicher Wörter hoch sein kann. Deshalb wird zunächst eine Neigungs- und Höhennormalisierung der segmentierten Wortabbilder durchgeführt. Außerdem kann für die zu suchenden Schlüsselwörter angenommen werden, dass jeweils mehrere Beispiele vorliegen, welche für die Anfrage genutzt werden dürfen. Der Anfragetyp wird deshalb als *Query-by-Wordclass* bezeichnet. Eine *Wortklasse* umfasst alle für ein Schlüsselwort zu Verfügung stehenden Anfragebeispiele (vgl. [RP09]).
Die vorgeschlagene sequentielle Wortrepräsentation nennen Rodriguez und Perronnin "local gradient histogram features". Sie benutzen ein gleitendes Fenster, um eine Sequenz von lokalen Bilddeskriptoren zu erzeugen. Diese lokalen Bilddeskriptoren sind jeweils dem, in Kapitel 2.3.1, erklärten SIFT-Deskriptor sehr ähnlich und werden im weiteren Verlauf dieser Arbeit als *LGH-Deskriptor* bezeichnet. Bei dem LGH-Deskriptor handelt es sich ebenfalls um konkatenierte, in einer Gitterstruktur angeordnete Gradientenhistogramme. Die wesentlichen Unterschiede zum SIFT-Deskriptor bestehen darin, dass bei dem LGH-Deskriptor keine Gewichtung der Bildgradienten vorgenommen wird (vgl. 2.3.1). Im Gegensatz zum SIFT-Deskriptor, bei dem eine trilineare Interpolation angewendet wird, wird beim LGH-Deskriptor eine lineare Interpolation zur Quantisierung der Bildgradienten verwendet. Für den LGH-Deskriptor wird außerdem eine einfachere Normalisierung vorgeschlagen. Die Komponenten werden so skaliert, dass sie in der Summe 1 ergeben. Damit können Kontrastunterschiede innerhalb eines Wortabbildes ausgeglichen werden (vgl. [RP08]).
Rodriguez und Perronnin schlagen drei verschiedene Varianten für die *local gradient*

histogram features vor.
Die erste Variante besteht darin, den LGH-Deskriptor über den gesamten Fensterausschnitt zu berechnen.
Deutlich bessere Ergebnisse lieferte die zweite Variante. Dabei wird der Deskriptor auf den im Fenster vorkommenden Schriftbereich angepasst [RP08]. Abbildung 3.1.1 zeigt das Schema der Extraktion der *local gradient histogram features* für die zuvor beschriebenen Varianten. Für diese sind alle Zellen eines LGH-Deskriptors jeweils gleich groß und werden gegenüber der dritten Variante als regulär bezeichnet (vgl. [RP08]).
Bei der dritten Variante wird die Gitterstruktur des LGH-Deskriptors angepasst, sodass der Bereich zwischen der oberen und der unteren Grundlinie feiner aufgelöst wird als der Bereich oberhalb bzw. unterhalb dieser Linien. Für die dritte Variante ergibt sich somit eine unregelmäßige Gitterstruktur. Diese Variante ist hier nur der Vollständigkeit halber aufgeführt und für den weiteren Verlauf der vorliegenden Arbeit nicht relevant. Dies hat folgende Gründe: Zum einen benötigt diese Anpassung eine robuste Schätzung für die obere und untere Grundlinie der segmentierten Wortabbilder. Zum anderen wurde bereits in [RP08] gezeigt, dass die zweite Variante zu besseren Ergebnissen führt.
Die Leistungssteigerung durch die Anpassung des LGH-Deskriptors auf den Schriftbereich wird durch die Beschränkung auf den Bereich mit dem größten Informationsgehalt erklärt (vgl. [RP08] Abschnitt 4.4).
Für den LGH-Deskriptor wurde eine Zellenaufteilung von 4×4 und eine Anzahl von $T = 8$ Hauptorientierungen als gute Konfiguration für den verwendeten Datensatz ermittelt. Dies führt, wie bei der für den SIFT-Deskriptor in [Low04] ermittelten optimalen Konfiguration, zu einem 128-dimensionalen Merkmalsvektor. Die Dimensionalität der vorgeschlagenen Wortrepräsentation ist somit relativ hoch (vgl. [RP08]).
Rodriguez und Perronnin haben ihre *local gradient histogram features* mit anderen in der Literatur zum Word Spotting oder zur Texterkennung vorgeschlagenen Merkmalen verglichen. Ihre Experimente zeigen, dass sich die von ihnen beschriebenen Merkmale auf den untersuchten Daten im untersuchten System besser zum Word Spotting eignen als die Anderen. In den Experimenten wurden unter anderem auch die zuvor erwähnten Merkmale aus [RM03] verglichen (vgl. [RP08] Abschnitt 3.3). Die Merkmale, welche in [RP08] am zweitbesten abgeschnitten haben, werden auch in der vorliegenden Arbeit näher betrachtet und in Abschnitt 3.1.2 eingeführt.
Zur Ähnlichkeitsbewertung haben Rodriguez und Perronnin Experimente mit Dynamic-Time-Warping (DTW) und Hidden-Markov-Modellen durchgeführt. Bei Hidden-Markov-Modellen (HMMs) handelt es sich um ein statistisches Verfahren, bei dem in einer vorherigen Trainingsphase spezifische Modelle geschätzt werden. In [RP08]

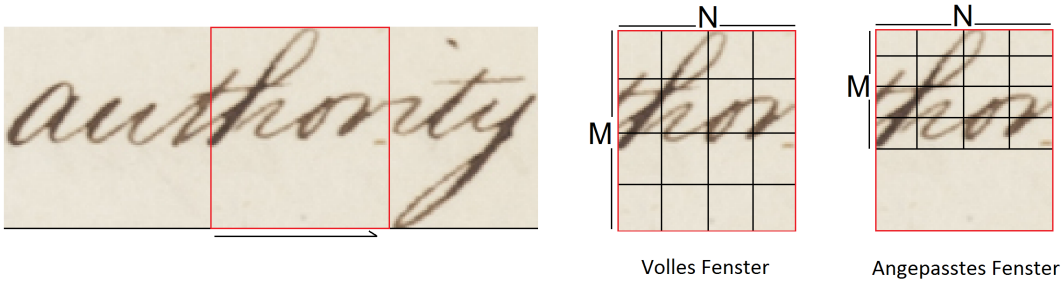

Abbildung 3.1.1: Extraktion der "local gradient histogram features" mit gleitendem Fenster. Abbildung nach [RP08]

werden auf diese Weise Modelle für die zu suchenden Schlüsselworte geschätzt. Als Trainingsbasis stehen dabei alle Wortabbilder einer Wortklasse zu Verfügung. Die geschätzten Modelle können anschließend für die Vergabe eines Ähnlichkeitswertes für jedes segmentierte Wortabbild genutzt werden. Mit den HMMs ließen sich in den Experimenten bessere Resultate für das beschriebene *Query-by-Wordclass* erzielen (vgl. [RP08]). Als Grund dafür wird hier die große Trainingsbasis gesehen. In einem QbE Word Spotting Szenario wären die Unterschiede zwischen DTW und HMMs vermutlich nicht so groß.
Die Auswertung der *local gradient histogram features* erfolgte für ein unbeschränktes Word Spotting Szenario auf einer Sammlung von eingescannten französischen Kundenbriefen eines Unternehmens. Nach bestem Wissen des Autors wurde dieser Datensatz nicht in anderen Arbeiten zur Auswertung von Word Spotting verwendet. Daher ist ein direkter Vergleich der vorgeschlagenen Methodik zu anderen Arbeiten nicht möglich.

3.1.2 *Vinciarelli Merkmale*

Bereits im Jahr 2000 haben Vinciarelli und Luettin in "Off-Line Cursive Script Recognition based on continuous density HMM" [AJ00] die gleiche Anpassungsmethode wie Rodriguez und Perronnin für ihre Extraktion einer Merkmalssequenz zur Schrifterkennung verwendet. Auch sie benutzen ein gleitendes Fenster. Ausgehend von einem Binärbild wird in jedem Fenster der Bereich, welcher Schriftpixel beinhaltet, in 4×4 gleich große Zellen unterteilt. Für jede dieser Zellen wird der Anteil der Schriftpixel innerhalb der Zelle relativ zur Anzahl der Schriftpixel im gesamten Fenster als Merkmal betrachtet [AJ00]. Prinzipiell beschreibt der Deskriptor somit die räumliche Verteilung der Schriftpixel des lokalen Bildausschnittes. Auf diese Weise entsteht

ein 16-dimensionaler Merkmalsvektor, welcher im weiteren Verlauf der Arbeit als VIN-Deskriptor bezeichnet wird. Um auch für diesen Deskriptor den Bezug zum Titel dieser Arbeit herzustellen, sei auf das Kapitel 4.3.4 von [Pri15] verwiesen. Demnach lässt sich dieser Deskriptor auch als ein räumliches Histogramm auffassen.
Obwohl in [RP08] beispielhaft gezeigt wird, dass der LGH-Deskriptor besser zum Word Spotting in einem Mehrschreiber Szenario geeignet ist, so lohnt sich die Betrachtung des VIN-Deskriptors dennoch aufgrund seiner Einfachheit und vergleichsweise geringen Dimensionalität.

3.2 WORD SPOTTING MIT BAG-OF-FEATURES

Das in Kapitel 2.3.2 vorgestellte Bag-of-Features Konzept wurde für Word Spotting zum ersten Mal von Rusiñol et al. [RATL11] vorgeschlagen. Seitdem sind weitere Vorschläge für Word Spotting Ansätze basierend auf Bag-of-Features gefolgt (z.B. [RRF13], [RRLF14], [SF15]). Im Folgenden wird zunächst die Arbeit von Rusiñol et al. vorgestellt. Anschließend werden die Arbeiten vorgestellt, welche die BoF-Sequenz dieser Arbeit motivieren. Zum Schluss dieses Unterkapitels wird auf weitere Arbeiten zum Word Spotting eingegangen, welche in Kapitel 5.4 für den Vergleich zum State-of-the-Art herangezogen werden.
Rusiñol et al. stellen in ihrem Paper "Browsing Heterogenous Document Collections by a Segmentation-free Word Spotting Method" [RATL11] eine auf SIFT-Deskriptoren aufbauende, Bag-of-Features basierte, holistische Wortrepräsentation für ein segmentierungsfreies QbE Word Spotting Szenario vor. Anstatt einer Segmentierung verwenden sie einen *Patch-basierten* Ansatz. Die Dokumente werden dabei jeweils in eine Menge von gleich großen Bildausschnitten (Patches) unterteilt, welche in einem dichten Gitter über das Dokument angeordnet sind. Für jeden Patch wird angenommen, dass er ein potentieller Wortkandidat ist, und eine Repräsentation berechnet. Zur Anfragezeit werden die Repräsentationen der Patches mit der des Anfragebildes verglichen. An Regionen, die der Anfrage ähneln, werden den Patches jeweils hohe Ähnlichkeitswerte zugewiesen. Die lokal optimalen Patches dieser ähnlichen Regionen werden als Antwortliste des Words Spotting zurückgegeben. Eine Segmentierung als Vorverarbeitungsschritt und damit mögliche unwiderrufbare Fehler können so vermieden werden [RATL11].
Für die in [RATL11] vorgeschlagene Wortrepräsentation werden zunächst SIFT-Deskriptoren in einem dichten Gitter berechnet. Pro Gitterpunkt werden Deskriptoren in drei verschiedenen Größen, welche sich an der Schriftgröße orientieren, berechnet. Dadurch soll gewährleistet werden, dass verschiedene Ausschnitte der Buchstaben eines Wortes

abgedeckt werden [RATL11]. Anschließend wird ein visuelles Vokabular mittels Clustering erzeugt und die SIFT-Deskriptoren auf dieses visuelle Vokabular quantisiert.
Anstelle eines einfachen Bag-of-Features Histogramms über den gesamten Bildauschnitt der jeweiligen Wortkandidaten werden diese auf mehreren Hierarchieebenen jeweils in gleich große, in einem Gitter angeordnete Zellen unterteilt. Für jede dieser Zellen wird dann ein Bag-of-Features Histogramm berechnet. Die Konkatenation der Bag-of-Features Histogramme ergibt die holistische Wortrepräsentation. Das Verfahren wird auch als *Spatial Pyramid* (SP) bezeichnet [LSP06]. Durch die SP können räumliche Informationen in der Repräsentation erfasst werden. Es hat sich gezeigt, dass es wichtig sein kann, räumliche Informationen in die Repräsentation für Wortkandidaten zu integrieren (vgl. [RRLF14]).
Ein Nachteil dieser Bag-of-Features basierten SP Repräsentation (SP-BoF) ist die hohe Dimensionalität. Diese ergibt sich durch die Anzahl der Zellen der SP und der Dimensionalität der Bag-of-Features Histogramme. Für den praktikablen Einsatz sind Verfahren zur Dimensionsreduktion nötig (vgl. [RATL11], [SF15]). Zur Ähnlichkeitsbewertung wird anschließend die Kosinus-Distanz (siehe 2.4) verwendet.

Bag-of-Features HMMs

2013 haben Rothacker et. al. in "Bag-of-Features HMMs for segmentation-free word spotting in handwritten documents" [RRF13] eine auf SIFT-Deskriptoren im dichten Gitter aufbauende, Bag-of-Features basierte, sequentielle Wortrepräsentation für ein segmentierungsfreies QbE Word Spotting Szenario vorgeschlagen. Die Wortrepräsentation setzt dabei auf dem zuvor beschriebenen Patch-basierten Ansatz aus [RATL11] auf. Ein Unterschied besteht allerdings darin, dass die SIFT-Deskriptoren nur in einer Größe berechnet werden (vgl. [RRF13]). Zur Integration räumlicher Informationen in die Repräsentation der Wortkandidaten wird eine Sequenz von Bag-of-Features Histogrammen vorgeschlagen. Diese sequentielle Repräsentation ermöglicht den Einsatz von HMMs zur Ähnlichkeitsbewertung. Obwohl als Trainingsbasis für die Modelle im QbE Szenario nur das Anfragewortabbild zu Verfügung steht, konnten die Word Spotting Ergebnisse auf dem Auswertungsdatensatz aus [RATL11] deutlich übertroffen werden (vgl. [RRF13]).
In [RRLF14] wird die Arbeit aus [RRF13] erweitert. Dabei wird vor allem einem Problem des Patch-basierten Ansatzes begegnet. Bei dem Patch-basierten Ansatz müssen sehr viele Wortkandidaten repräsentiert werden. Durch die simple Anordnung der Patches im dichten Gitter werden sehr viele Wortkandidaten betrachtet, möglicherweise auch Solche, welche kein Wort enthalten. In [RRLF14] wird diese simple Anordnung

der Patches in einem dichten Gitter umgangen, indem für jede Anfrage eine Voranalyse der Dokumente durchgeführt wird. Auf diese Weise lassen sich die Regionen finden, bei denen sich eine detailliertere Betrachtung lohnt. Der Rechenaufwand kann somit reduziert werden.
Ein anderes Problem des Patch-basierten Ansatzes ist die Größenwahl der Patches. In [RATL11] wird eine einheitliche Größe unabhängig von der Anfrage gewählt. Die Größenwahl ist abhängig von der Auflösung der Dokumente und der Größe der Schrift. Eine passende Größe muss also für jeden Datensatz neu ermittelt werden. Patches, welche kleine Worte enthalten, erfassen neben dem Wort auch andere Teile des Dokumentes, die irrelevant für das Wort sind [RRLF14]. In [RRLF14] entspricht die Größe der Wortkandidaten der des Anfragebildes. Diese Größenwahl ist naheliegend, aber nur gerechtfertigt, solange die Größenvariation der Schrift innerhalb der betrachteten Dokumente nicht zu groß ist. Gerade bei einem Datensatz mit Dokumenten mehrerer Schreiber kann dies der Fall sein. Bei Verfahren zur Segmentierung werden diese Größenvariationen in der Regel implizit berücksichtigt (z.B. [Man99]).
Die Diskussion über den Patch-basierten Ansatz bzw. die Vor- und Nachteile zwischen segmentierungsfreiem und segmentierungsbasiertem Word Spotting soll an dieser Stelle nicht weiter vertieft werden. Denn für beide Ansätze ist die Repräsentation der Wortkandidaten ein wichtiger Schritt. Zudem lassen sich moderne Wortrepräsentationen häufig in beiden Szenarien einsetzen (siehe [SRF15], [PTV15]).

Weitere Arbeiten

2015 haben Sudholt und Fink in [SF15] ein Verfahren zur Dimensionsreduktion für eine SP-BoF Wortrepräsentation in einem segmentierungsbasierten QbE Word Spotting Szenario vorgeschlagen. Diese SP-BoF Repräsentation basiert ebenfalls auf SIFT-Deskriptoren im dichten Gitter. Ohne auf Details zu dem Verfahren der Dimensionsreduktion einzugehen sei hier eine Grundlage für den Erfolg des Verfahrens angeführt: Die BrayCurtis-Distanz (siehe 2.4) eignet sich besonders für den Vergleich der verwendeten SP-BoF Repräsentation [SF15]. Dieser Ansatz ohne Dimensionsreduktion wurde auch bei der "Competition on Keyword Spotting for Handwritten Documents" auf der ICDAR 2015 von der *Pattern Recognition Group* (PRG) der TU Dortmund eingereicht. Die Autoren von [SF15] gehören zur PRG. Mit diesem Ansatz erreichte man die besten Ergebnisse bei der "Competition" (vgl. [PTV15]).
Alle bisher vorgestellten Arbeiten zum Word Spotting mit Bag-of-Features haben gemeinsam, dass sie auf dem SIFT-Deskriptor basieren. Neben dem SIFT-Deskriptor finden sich nur wenige andere lokale Bilddeskriptoren als Grundlage für das vi-

suelle Vokabular beim Bag-of-Features basierten Word Spotting. In [ARTL15] wird beispielsweise eine Variante des HOG-Deskriptors eingesetzt. Dieser Deskriptor wurde hauptsächlich aus Laufzeitgründen vorgeschlagen (siehe [ARTL15] Abschnitt 2.2). Bei der Wortrepräsentation in [ARTL15] handelt es sich ebenfalls um eine SP-BoF Repräsentation. Diese wird allerdings zusätzlich, durch eine Reihe von Vorschlägen zur Optimierung, verändert. Für die Ähnlichkeitsbewertung wird die L2-Distanz verwendet. Auch die Autoren von [ARTL15] nahmen an der "Competition" auf der ICDAR 2015 als Gruppe "Computer Vision Center (CVC)" teil.
Die zuletzt vorgestellten Methoden wurden für ein segmentierungsbasiertes Szenario ausgewertet und dienen daher für den späteren Vergleich zum State-of-the-Art (siehe Kapitel 5.4).

ZUSAMMENFASSUNG

Aus [RM03] stammt die Idee zum DTW der vorliegenden Arbeit. Das DTW wird auch in der vorliegenden Arbeit mit dem Ziel eingesetzt, Schriftvariationen in Schreibrichtung kompensieren zu können. In [RP08] wurde die Wortrepräsentation der *local gradient histogram features* vorgestellt. Es wurde festgestellt, dass es sich bei dieser Repräsentation um eine Sequenz von LGH-Deskriptoren handelt, welche dem SIFT-Deskriptor ähneln (siehe 3.1.1). In einer solchen D-Sequenz lassen sich auch andere Deskriptoren wie der VIN-Deskriptor (siehe 3.1.2) einsetzen. Weiterhin wurde für die *local gradient histogram features* festgestellt, dass die Anpassung des LGH-Deskriptors auf den Schriftbereich zu einer Verbesserung des Word Spotting führen kann. Die veröffentlichten Auswertungsergebnisse beziehen sich allerdings auf einen in der Literatur nicht wiederverwendeten Datensatz. Ein Vergleich zu moderneren Ansätzen zum Word Spotting, speziell zu den Bag-of-Features Ansätzen, und damit auch die Beantwortung der Eingangsfrage, ist an dieser Stelle also noch nicht möglich.
Im zweiten Abschnitt des Kapitels (3.2) wurden einige Bag-of-Features basierte Ar-

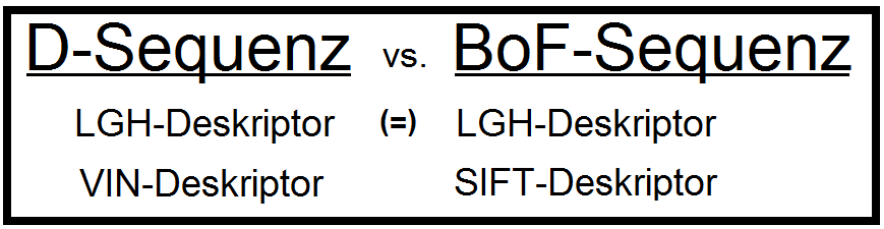

Abbildung 3.2.1: Übersicht zu den Untersuchungsaspekten dieser Arbeit.

beiten zum Word Spotting vorgestellt. In [RATL11] wurde ein Bag-of-Features Ansatz zum ersten Mal für segmentierungsfreies Word Spotting vorgeschlagen. Die Idee zur BoF-Sequenz, in der Form, wie sie in der vorliegenden Arbeit eingesetzt wird, stammt aus [RRF13]. Diese wurde ebenfalls in einem segmentierungsfreien Word Spotting Szenario untersucht. Als Grundlage für die Bag-of-Features der vorgestellten Arbeiten wird meistens der SIFT-Deskriptor im dichten Gitter eingesetzt. Es können aber auch andere lokale Bilddeskriptoren eingesetzt werden (z.B. [ARTL15]). Für den Vergleich zwischen BoF-Sequenz und D-Sequenz, und damit zur Beantwortung der Eingangsfrage, wird hier der LGH-Deskriptor verwendet.
Die Übersicht in Abbildung 3.2.1 fasst die Untersuchungsaspekte der vorliegenden Arbeit zusammen. Diese Übersicht leitet gleichzeitig das Kapitel 4 dieser Arbeit ein.

4 METHODIK

In diesem Kapitel wird die Methodik der vorliegenden Arbeit erläutert. Die Methodik verfolgt hier zwei Ziele: Eines ist das erfolgreiche Word Spotting auf historischen Dokumenten. Das Andere ist der Vergleich zwischen der D-Sequenz und der BoF-Sequenz. Dazu werden die beiden Ansätze für sequentielle Wortrepräsentationen jeweils in Kombination mit verschiedenen lokalen Bilddeskriptoren untersucht. Abbildung 3.2.1 sei hier noch einmal zur Übersicht referenziert.
Zunächst wird das Verfahren der D-Sequenz erklärt. Danach folgt die Erläuterung der BoF-Sequenz. Anschließend wird auf die lokalen Bilddeskriptoren eingegangen, welche in beiden Verfahren eingesetzt werden. Vor der Zusammenfassung des Kapitels wird der DTW-Algorithmus erklärt, welcher hier für die Ähnlichkeitsbewertung verwendet wird. Bei den Erläuterung werden die jeweiligen Parameter und ihre Eigenschaften betrachtet.
Analog zu [ZPG14] und [SF15] wird hier für die Methodik angenommen, dass eine Segmentierung der Dokumente vorliegt.

4.1 DESKRIPTOR SEQUENZ

Die Methodik der Deskriptor Sequenzen (D-Sequenzen) orientiert sich im Wesentlichen an der Methodik, welche in [RP08] zur Repräsentation von Wortkandidaten beschrieben wird (siehe auch Kapitel 3.1.1).
Die D-Sequenzen werden mit einem gleitenden Fenster extrahiert. Auch hier lässt sich der Detailgrad der Wortrepräsentation durch die *Schrittweite des Fensters* variieren. Wie in [RP08] werden hier zwei Varianten vorgeschlagen. Bei der einfacheren Variante wird der Deskriptor für den gesamten Fensterbereich berechnet. Es lässt sich vermuten, dass sich mit dieser Variante bei wenig Schriftvariation und einer guten Segmentierung gute Word Spotting Ergebnisse erzielen lassen.
Bei der anderen Variante wird der Deskriptor auf den Bereich, welcher Schriftpixel enthält, angepasst (siehe 3.1.1). Sowohl für die angepasste Variante als auch für die einfache Variante lässt sich durch die *Fensterbreite* des gleitenden Fensters der Bildausschnitt, welchen der jeweilige Deskriptor erfasst, einstellen. Um in der angepassten Variante den jeweiligen Bereich der Schriftpixel zu identifizieren, wird eine

Binarisierung durchgeführt. In dieser Arbeit wird eine einfache Binarisierung auf Basis des Niblack-Schwellwertes durchgeführt (siehe Kapitel 2.1). Mit dem Parameter α kann die Qualität der Binarisierung verändert werden. Wird der Parameterwert betragsmäßig zu groß eingestellt, so werden nicht alle Schriftpixel erfasst. Wird er zu klein eingestellt, so wird der Hintergrund als Schrift deklariert (siehe Abbildung 2.1.1). Der Schwellwert wird jeweils für den gesamten Bildbereich eines segmentierten Wortabbildes berechnet und angewendet. Diese Methode ist einfach und führt auf den betrachteten Datensätzen zu guten Ergebnissen (siehe Auswertung 5.2.1).
In dem binarisierten Wortabbild lassen sich die obere und untere Wortkontur bestimmen. Diese grenzen den Bereich der Schriftpixel ein. Abbildung 4.1.1 zeigt das Schema für diese Anpassung anhand eines Beispiels. Innerhalb des gleitenden Fensters wird der Deskriptor auf den Bildbereich zwischen dem höchsten Punkt der oberen und dem tiefsten Punkt der unteren Wortkontur angepasst. Ein wesentlicher Unterschied zu [RP08] ergibt sich in dieser Arbeit dadurch, dass keine Neigungsnormalisierung durchgeführt wird. Diese wird hier bewusst weggelassen, um dem Szenario historischer Dokumente gerecht zu werden, bei dem solche Vorverarbeitungsschritte nicht immer einfach sind.
Wie in Abbildung 3.2.1 zu sehen ist, werden auch für die D-Sequenz mehrere Varianten vorgeschlagen. Für den LGH-Deskriptor werden sowohl die Variante ohne Anpassung, im Folgenden durch LGH-Full abgekürzt, als auch die angepasste Variante untersucht. Letztere wird aufgrund der beschriebenen Binarisierung mit LGH-Niblack abgekürzt. Der VIN-Deskriptor wird in der D-Sequenz, wie ursprünglich vorgeschlagen (siehe Kapitel 3.1.2), in der angepassten Variante untersucht (VIN-Niblack).

4.2 BAG-OF-FEATURES SEQUENZ

Die Idee für die Bof-Sequenz stammt aus [RRLF14] und vorherigen Arbeiten [RRF13], [RVF12].
Für die Erstellung eines visuellen Vokabulars wird zunächst eine geeignet große Stichprobe lokaler Bilddeskriptoren benötigt (siehe auch [RRLF14]). Dazu werden zufällig segmentierte Wortabbilder ausgewählt. Nach Möglichkeit werden alle Wortabbilder verwendet, da so alle Charakteristiken der Daten erfasst werden können. Anschließend werden lokale Bilddeskriptoren einer Größe in einem dichten Gitter berechnet. Wie in [RRLF14] werden die Deskriptoren via Lloyd-Clustering (siehe Kapitel 2.3.2) zu einem visuellen Vokabular zusammengefasst. Die *Größe des visuellen Vokabulars* ist ein Parameter der Methodik. Je größer das visuelle Vokabular, desto spezifischer werden die einzelnen visuellen Worte und damit auch die Bag-of-Features Histogramme.

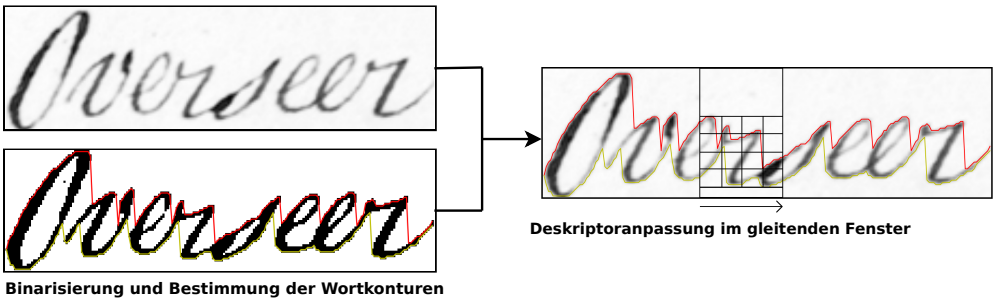

Abbildung 4.1.1: Schema zur Extraktion der angepassten D-Sequenzen anhand eines Beispiels. Links oben: Original als Grauwertbild. Links unten: Das binarisierte Bild mit oberer (rot) und unterer (gelb) Wortkontur. Rechts: Die Anpassung des Deskriptors auf den Schriftbereich. (Abbildung in Teilen nach [Tru07].)

Die BoF-Sequenz eines Wortabbildes ergibt sich wie folgt: Auf dem Wortabbild werden lokale Bilddeskriptoren einer Größe in einem dichten Gitter berechnet. Die Abstände in beide Richtungen der Bildmatrix zwischen den einzelnen Deskriptoren sind die zwei Parameter des dichten Gitters. Je kleiner diese Abstände sind, desto größer wird die *Auflösung des Gitters*. Es werden nur die Deskriptoren betrachtet, welche vollständig innerhalb des Wortabbildes liegen. So müssen keine Annahmen zu Erweiterungen am Rand des Wortabbildes getroffen werden. Die *horizontale und vertikale Größe der Deskriptoren* sind weitere Parameter. Die lokalen Bilddeskriptoren, ihre Parameter und die Parameter des dichten Gitters haben dabei die gleichen Werte wie bei der Erstellung des visuellen Vokabulars. Es werden keine Deskriptoren, wie beispielsweise in [RATL11], verworfen.
Nach der Berechnung der Deskriptoren werden diese durch eine nächste Nachbarzuordnung auf das visuelle Vokabular quantisiert. Anschließend wird ein gleitendes Fenster über das so entstehende Gitter von charakteristischen Merkmalen bewegt. Der horizontale Abstand der Deskriptoren im Deskriptorgitter bestimmt hier auch die minimale *Schrittweite des Fensters* in Pixeln. Je höher die Gitterauflösung und je kleiner die Fensterschrittweite, desto detaillierter wird das Wortabbild beschrieben. Die *Fensterbreite* beschränkt sich in dieser Arbeit wie in [RRLF14] auf die Breite der einzelnen Deskriptoren, also auf eine Spalte im Merkmalsgitter. An jeder Fensterposition wird ein Bag-of-Features Histogramm über die Merkmale innerhalb des Fensters gebildet. In Analogie zu [RRF13] werden die Histogramme anschließend normalisiert. Hier werden die Histogramme so normalisiert, dass die Summe der Komponenten 1 ergibt. In Abbildung 4.2.1 ist der zuvor beschriebene Ablauf zur Berechnung für

die BoF-Sequenzen visualisiert. Dort lässt sich auch erkennen, dass die einzelnen Bag-of-Features Histogramme bei großen visuellen Vokabularen in der Regel sehr dünn besetzt sind, also viele 0-Einträge besitzen.
Falls ein Wortabbild kleiner ist als der Deskriptor, wird das Wortabbild vorher auf die minimale Größe skaliert, die nötig ist, um mindestens einen Deskriptor berechnen zu können. Dabei wird das Seitenverhältnis beibehalten.
Wie in Abbildung 3.2.1 werden in dieser Arbeit zwei Varianten der BoF-Sequenz betrachtet. Bei der einen wird der SIFT-Deskriptor als Grundlage für das visuelle Vokabular und die charakteristischen Merkmale verwendet (SIFT-BoF). Die SIFT-BoF Variante dient in dieser Arbeit als Referenz zum State-of-the-Art.
Bei der anderen Variante wird der LGH-Deskriptor anstelle des SIFT-Deskriptors eingesetzt (LGH-BoF). Die LGH-BoF soll die Vergleichbarkeit zur D-Sequenz gewährleisten (siehe auch Abbildung 3.2.1).
Im folgenden Abschnitt wird wiederholt auf einige Details und Parameter der betrachteten lokalen Bilddeskriptoren eingegangen.

4.3 LOKALE BILDDESKRIPTOREN

Basis für die vorgeschlagenen sequentiellen Repräsentationen sind jeweils lokale Bilddeskriptoren. Bei der BoF-Sequenz dienen sie als Grundlage für die Bildung charakteristischer Merkmale, dem visuellen Vokabular. Bei der D-Sequenz werden sie direkt als Merkmalsvektoren interpretiert (siehe auch Kapitel 2.3). Neben dem SIFT-Deskriptor, welcher bereits in Kapitel 2.3.1 vorgestellt wurde, werden hier die folgenden Deskriptoren in mindestens einer der Sequenzen eingesetzt (siehe Abbildung 3.2.1).

VIN-Deskriptor

Bei dem VIN-Deskriptor handelt es sich um den einfachsten der hier betrachten Deskriptoren, sowohl im Bezug auf Berechnungskomplexität als auch auf die Dimensionalität (siehe Kapitel 3.1.2). Die Einteilung des Deskriptors in $M \times N$ Zellen ist die einzige Einstellungsmöglichkeit des VIN-Deskriptors. Durch die Variation der Anzahl der Zeilen (M) bzw. Spalten (N) lässt sich der Detailgrad der räumlichen Beschreibung in vertikaler bzw. horizontaler Richtung einstellen.

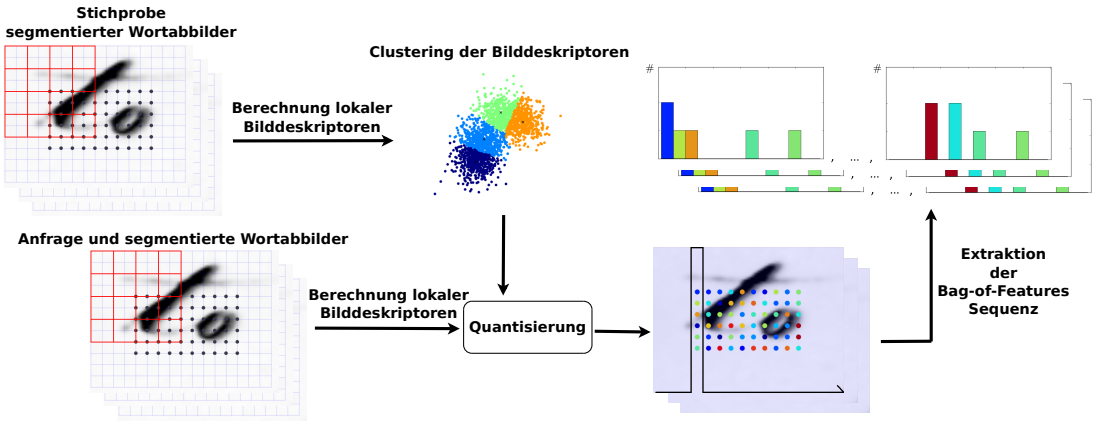

Abbildung 4.2.1: Schema zur Berechnung der BoF-Sequenzen. (In Teilen nach [RRLF14] Fig. 2)

LGH-Deskriptor

In dieser Arbeit wird die reguläre Variante des LGH-Deskriptors verwendet, wie sie in [RP08] vorgeschlagen wird. Als Grundlage für die Gradienten wird hier der Sobel-Operator (2.2) verwendet, um eine Glättung in den Schritt der Gradientenberechnung zu integrieren.
Die *Einteilung des Deskriptors* in $M \times N$ Zellen und die *Anzahl der Hauptorientierungen* der Gradientenhistogramme sind die Einstellungsmöglichkeiten des LGH-Deskriptors. Wie bei dem VIN-Deskriptor lässt sich durch die Variation von M und N der Detailgrad der räumlichen Beschreibung des LGH-Deskriptors einstellen. Durch die Variation der Anzahl der Hauptorientierungen der Gradientenhistogramme lässt sich zusätzlich der Detailgrad der Beschreibung des Schriftverlaufs erhöhen.
Aufgrund der in Kapitel 3.1.1 festgestellten Ähnlichkeiten ergeben sich die gleichen Parameter für den SIFT-Deskriptor. An dieser Stelle wird er daher nicht explizit aufgeführt.

4.4 DYNAMIC-TIME-WARPING

Der DTW-Algorithmus dieser Arbeit orientiert sich an dem DTW-Algorithmus, welcher in [RM03] vorgeschlagen und auch in [RP08] verwendet wird. Als Steilheitseinschränkung wird die in der Literatur als "local continuity constraint" bezeichnete Einschränkung verwendet [RM03]. Für ein Tupel C_{k+1} der Warping Funktion müssen mindestens einer und maximal beide Einträge direkte Nachfolger in den jeweiligen

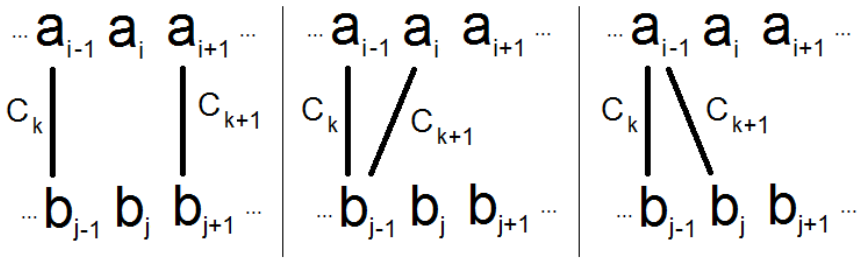

Abbildung 4.4.1: Schema möglicher Zuordnungen zweier aufeinander folgender Zuordnungspaare der Warping Funktion beim "local continuity constraint".

Merkmalssequenzen der Einträge von C_k sein. Abbildung 4.4.1 zeigt zur Verdeutlichung alle möglichen Nachfolger von C_k unter Berücksichtigung dieser Steilheitseinschränkung.
Für diese Arbeit ergibt sich somit die folgende Berechnungsvorschrift für den DTW-Algorithmus:

$$D(i,j) = \min\{D(i-1,j), D(i,j-1), D(i-1,j-1)\} + d(a_i, b_j) \quad \text{nach [RM03]} \quad (4.4.1)$$

Da das *Distanzmaß* $d(a_i, b_j)$ entscheidend für den Vergleich von Merkmalsvektoren ist, werden bei der Auswertung verschiedene Distanzmaße berücksichtigt (siehe Kapitel 2.4).
Mit der Berechnungsvorschrift lässt sich ein Feld nach dem Konzept der dynamischen Programmierung füllen. Es werden dabei nur die Bereiche innerhalb des Anpassungsfensters berücksichtigt. Als Anpassungsfenster wird das in Kapitel 2.4 erwähnte "Sakoe-Chiba-Band" verwendet, um die Laufzeit zu reduzieren. Außerdem zeigt die Auswertung, dass das Anpassungsfenster die Qualität der Word Spotting Ergebnisse entscheidend beeinflussen kann (siehe Kapitel 5.4.1).
Die *Breite des Anpassungsfensters* begrenzt die Warping Funktion so, dass ausgehend von der diagonalen Warping Funktion eine maximale festgelegte Abweichung erlaubt wird. Abbildung 4.4.2 zeigt ein Beispiel für ein solches Array, welches mit dem DTW-Algorithmus dieser Arbeit berechnet wurde. Die Breite des Anpassungsfensters ist dabei auf 10% der längeren Sequenz, in diesem Fall also 2, festgelegt. Die Distanz zwischen den beiden Merkmalssequenzen ist rot hinterlegt. Durch Backtracking kann der Warping Pfad [RM03], also die Warping Funktion, welche zu dieser Distanz führt, bestimmt werden. Im Beispiel ist der Warping Pfad grau hinterlegt.
Um den Einfluss unterschiedlich langer Merkmalssequenzen der Wortkandidaten auf diese Distanz auszugleichen, wird die Distanz jeweils durch die Länge des Warping Pfades normalisiert (siehe auch [RM03]). Diese normalisierte Distanz zwischen

a_{10}	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	11,09	11,46	12	12,6
a_9	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	9,73	10,44	10,8	11,29	11,97	12,68
a_8	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	8,46	9,13	9,87	10,11	10,68	11,36	12,03	12,73
a_7	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	7,6	7,98	8,64	9,41	9,53	10,16	10,83	11,6	12,36	13,13
a_6	∞	∞	∞	∞	∞	∞	∞	∞	6,3	6,84	7,41	8,04	8,7	9	9,68	10,29	11,05	11,82	∞	∞
a_5	∞	∞	∞	∞	∞	∞	5,03	5,65	6,18	6,88	7,36	7,97	8,44	9,15	9,75	10,45	∞	∞	∞	∞
a_4	∞	∞	∞	∞	∞	3,67	4,32	4,97	5,48	6,07	6,62	7,25	7,83	8,52	9,12	∞	∞	∞	∞	∞
a_3	∞	∞	∞	2,67	3,03	3,7	4,33	4,83	5,34	5,96	6,64	7,19	7,88	∞	∞	∞	∞	∞	∞	∞
a_2	1,4	1,39	1,91	2,41	3,07	3,72	4,22	4,71	5,32	5,92	6,5	∞	∞	∞	∞	∞	∞	∞	∞	∞
a_1	0,63	1,25	1,85	2,5	3,16	3,67	4,18	4,77	5,34	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
A/B	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}	b_{14}	b_{15}	b_{16}	b_{17}	b_{18}	b_{19}	b_{20}

Abbildung 4.4.2: Beispiel eines DP-Array, nach der DTW Berechnung. Nach dem Algorithmus dieser Arbeit.

Anfrage und Wortkandidaten wird dann als Ähnlichkeitswert für die Antwortliste verwendet.
An dieser Stelle sei ein Nachteil des Dynamic-Time-Warping angeführt: Die Laufzeit des DTW-Algorithmus ist relativ hoch. Zur Berechnung der Distanz zwischen zwei Merkmalssequenzen der Länge I bzw. J muss ein Array der der Größe $I \times J$ gefüllt werden, die Laufzeit ist also durch $O(IJ)$ beschränkt.
In der Literatur wird der hohen Laufzeit oft mit sogenanntem *Pruning* entgegen gewirkt (z.B. [RM07], [RP09]). Dabei werden unwahrscheinliche Wortkandidaten aufgrund von einfachen heuristischen Entscheidungen verworfen [RM07]. So lässt sich der Berechnungsaufwand reduzieren. Es muss aber beachtet werden, dass durch das Pruning auch Wortkandidaten verworfen werden können, welche zur Anfrage passen [RM03]. Das kann die Anforderung der Vollständigkeit (siehe Kapitel 1.1) negativ beeinflussen. Daher wird hier kein Pruning vorgenommen.
Um die Auswertung hier in akzeptabler Zeit durchführen zu können, wurde die Word Spotting Implementierung für die Verteilung auf mehrere Rechner ausgelegt. Beim Word Spotting bietet sich eine solche Verteilung an. So wird auf jedem Rechner nur ein Teil aller Dokumente nach dem Anfragewort durchsucht. Für die reale Anwendung auf großen Datensätzen könnte eine solche Parallelisierung ebenfalls sinnvoll sein.

ZUSAMMENFASSUNG

In diesem Kapitel wurden die zu Beginn der Arbeit (siehe Kapitel 1.2) genannten Ansätze der D-Sequenz und BoF-Sequenz konkretisiert. Aus der jeweiligen Kombination der beiden Ansätze mit verschiedenen lokalen Bilddeskriptoren ergeben sich

nach Abbildung 3.2.1 die folgenden sequentiellen Wortrepräsentationen, welche im Evaluierungskapitel 5 untersucht werden: SIFT-BoF, LGH-BoF, LGH-Full, LGH-Niblack und VIN-Niblack.
Zur Übersicht sei hier noch einmal der Gesamtablauf des Word Spotting dieser Arbeit zusammengefasst. Mit Referenz auf die Kapitel 1.1 und 1.2 wird hier bezüglich des erste Schrittes, der Suche nach Wortkandidaten, ein segmentierungsbasiertes Word Spotting Szenario betrachtet. Die Segmentierung ist hier vorgegeben. Die zuvor benannten Methoden für sequentielle Wortrepräsentationen werden hier jeweils für den zweite Schritt der Repräsentation von Wortkandidaten eingesetzt. Zur Ähnlichkeitsbewertung, dem dritten Schritt, wird der beschriebene DTW-Algorithmus eingesetzt.
Eine Übersicht aller einstellbaren Parameter der Methodik ist im Anhang A hinterlegt.

5 EVALUIERUNG

In diesem Kapitel wird die vorgeschlagene Methodik hinsichtlich ihrer Eignung zum Word Spotting in historischen Dokumenten untersucht. In Kapitel 5.1 werden die Datensätze und Evaluierungsprotokolle für die Auswertung erläutert. In Kapitel 5.2 werden die BoF-Sequenz und die D-Sequenz, zunächst für sich betrachtet, ausgewertet. Die jeweiligen Varianten, welche in Kapitel 4.4 aufgezählt wurden, werden dabei nebeneinander gehalten. Es wird auf wichtige Parameter und deren Einflüsse auf die Qualität der Word Spotting Ergebnisse eingegangen. In Kapitel 5.3 erfolgt der Vergleich zwischen BoF-Sequenz und D-Sequenz. Zum Schluss erfolgt ein Vergleich zum State-of-the-Art in Kapitel 5.4. In diesem Zusammenhang wird auch auf einige Eigenschaften und Parameter des DTW-Algorithmus eingegangen.

5.1 DATENSÄTZE UND EVALUIERUNGSPROTOKOLLE

Für die Auswertung wurden exemplarisch zwei Datensätze historischer Handschriften ausgewählt, welche bereits in anderen Arbeiten für Auswertungen zum Word Spotting verwendet wurden. Für beide Datensätze steht jeweils eine *Ground Truth* zu Verfügung, die zur Beurteilung der Qualität der Word Spotting Ergebnisse genutzt werden. Die Auswahl mehrerer Datensätze ermöglicht eine differenzierte Sicht auf die Methodik. Besondere Schwierigkeiten der einzelnen Datensätze und Evaluierungsprotokolle können durch einen Vergleich besser erkannt werden. Im folgenden Teil dieses Unterkapitels werden die wichtigsten Details zu den Datensätzen aufgeführt und anschließend das verwendete Evaluierungsmaß erklärt.

George Washington

Bei dem ersten Datensatz handelt es sich um einen Auszug von 20 Seiten aus der "George Washington Papers collection"[1] der Library of Congress in Washington. Im Folgenden wird dieser Auszug als *GW Datensatz* bezeichnet. Die Schriftvariation innerhalb der Seiten des GW Datensatzes ist nicht besonders groß, deshalb wird die Auswertung auf diesem Datensatz häufig als Einzelschreiber Szenario gesehen

1 https://memory.loc.gov/ammem/gwhtml/gwhome.html

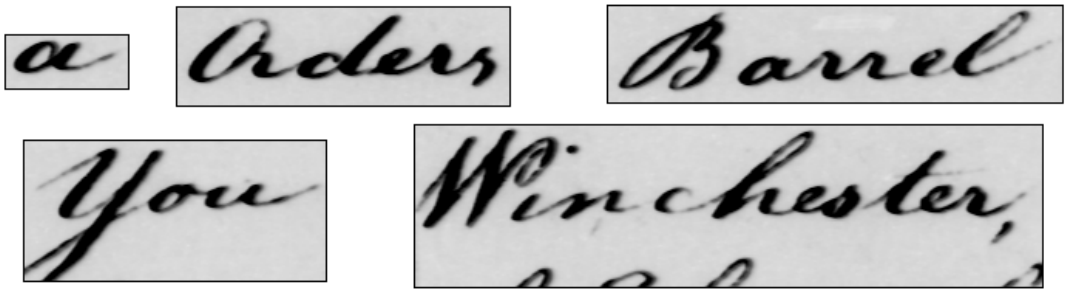

Abbildung 5.1.1: Beispiele von segmentierten Wortabbildern des GW Datensatzes. Eine Schwierigkeit: Beim Wort "Winchster" ragen Teile eines anderen Wortes mit ins Bild.

[RRLF14]. Die Ground Truth für die Dokumente enthält 4860 segmentierte Wortabbilder und wurde in [RM07] erstellt. Abbildung 5.1.1 zeigt einige Beispiele.
Die Dokumente sind auch für andere Arbeiten zur Auswertung verwendet worden (z.B. [SF15], [AFV13], [ZPG14]). Allerdings wurde nicht immer das gleiche Evaluierungsprotokoll verwendet, sodass ein direkter Vergleich erschwert wird. In dieser Arbeit wird nach dem Evaluierungsprotokoll aus [SF15] verfahren. Jedes Wort, welches mindestens zweimal in den Dokumenten vorkommt, wird als Anfrage verwendet. Dadurch ergeben sich 4221 Anfrageworte, welche zwischen einem und fünfzehn Zeichen lang sind.

Jeremy Bentham

Bei dem zweiten Datensatz handelt es sich um die "Bentham Collection"[2] des University College London (UCL). Die Bentham Collection enthält Schriften von dem Philosophen Jeremy Bentham sowie seinen Sekretären [PTV15] und ist Teil des im Kapitel 1 erwähnten Transcriptorium Projektes. Auszüge dieser Sammlung sind in der aktuellen Literatur für Auswertungen zum Word Spotting zu finden (z.B. [SF15], [ZPG14]). Durch den Wettbewerb zum Word Spotting auf der ICDAR 2015 sind zwei Auszüge aus der Bentham Collection samt Ground Truth und Evaluierungsprotokoll vorgegeben [PTV15]. Bei dem einen Auszug handelt es sich um den Evaluierungsdatensatz (JB-E), der für die finale Auswertung gedacht ist. Bei dem anderen Auszug handelt es sich um einen kleineren Validierungsdatensatz (JB-V), welcher zur Parame-

2 http://www.ucl.ac.uk/library/bentham

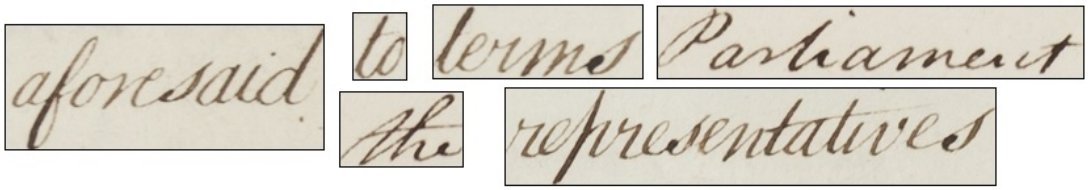

Abbildung 5.1.2: Beispiele segmentierter Wortabbilder des JB-V Datensatzes.

teroptimierung der Methodik verwendet werden kann. Beide Datensätze enthalten Dokumente verschiedener Schreiber. Die Schriftvariation ist daher höher als die des GW Datensatzes. Für die Parameterauswertung der Methodik wird hier der JB-V Datensatz verwendet. Dieser besteht aus 3234 segmentierten Wortabbildern, die aus insgesamt 10 Dokumentenabbildern der Bentham Collection stammen. Zusätzlich zu den segmentierten Wortabbildern sind 95 separate Anfragebilder von 20 unterschiedlichen Wörtern mit einer Länge von jeweils sechs oder mehr Zeichen vorgegeben. Abbildung 5.1.2 zeigt einige Beispiele.
Für einen späteren Vergleich zum State-of-the-Art wird die Methodik hier anschließend mit dem JB-E Datensatz des Wettbewerbs ausgewertet. Der Evaluierungsdatensatz besteht aus 15419 segmentierten Wortabbildern von 70 verschiedenen Dokumentenabbildern. Dazu sind 1421 Anfragebilder von 243 unterschiedlichen Worten, ebenfalls mit einer Länge von jeweils sechs oder mehr Zeichen, vorgegeben.

Evaluierungsmaß

In der vorliegenden Arbeit wird die *mean Average Precision* (mAP) als Evaluierungsmaß zur Beurteilung der Qualität der Word Spotting Resultate verwendet. Die mAP ist das arithmetische Mittel über die *Average Precision* (AP) aller Anfragen. Die AP einer Anfrage ist ein Maß dafür, wie gut die Antwortliste sortiert ist. Sie lässt sich wie folgt definieren (siehe auch [SF15] und [PTV15]):

$$AP = \frac{\sum_{k=1}^{n} (\pi(k) \times rel(k))}{|R|} \qquad \pi(k) = \frac{|R \cap S(k)|}{|R|} \tag{5.1.1}$$

Dabei ist R die Menge aller Wortabbilder aus der Retrieval-Datenbank, welche für die Anfrage relevant sind. $S(k)$ ist die Menge der ersten k Wortabbilder der Antwortliste zur Anfrage. $\pi(k)$ wird *Precision* an Position k genannt. $rel(k)$ ist eine Indikatorfunktion, welche 1 beträgt, wenn das Wortabbild an der Position k relevant. Sonst ist

$\mathrm{rel}(k) = 0$ [SF15]. Es sei an dieser Stelle gesagt, dass bei der Berechnung der AP für Anfrageergebnisse auf dem GW Datensatz das Anfragewortabbild nicht in R enthalten ist und aus der Antwortliste entfernt wird. Trivialerweise würde dieses immer an erster Position der Antwortliste stehen, sodass die Ergebnisse verzerrt würden. Die mAP wird hier immer in Prozent angegeben.

5.2 EXPERIMENTE UND AUSWERTUNG

Im Folgenden werden die Ergebnisse der Auswertungen vorgestellt. Die Auswertung der BoF-Sequenzen und der D-Sequenzen erfolgt jeweils in zwei Stufen. Zunächst werden *Basisexperimente* zu einer ersten generellen Einschätzung der Wortrepräsentationen und dem Verhalten des DTW durchgeführt. Das Hauptziel dieser Experimente ist, ein geeignetes Distanzmaß für das DTW (siehe Kapitel 2.4) zu finden.
Anschließend folgen Experimente zu den Parametern der jeweiligen Merkmale. Ziel ist hierbei das Finden einer möglichst guten Parameterkonfiguration und die Untersuchung der Einflüsse der jeweiligen Parameter.

5.2.1 *Basisexperimente*

Für die Basisexperimente wurden die folgenden Parameter eingestellt. Aus Laufzeitgründen wurde die Breite des Anpassungsfensters des DTW für die weiteren Experimente auf 10% der jeweils längeren Sequenz eingestellt. Die Schrittweiten der gleitenden Fenster orientierten sich jeweils an den Abständen, welche in [SF15] für das Deskriptor-Gitter verwendet wurden. Die gleichen Abstände wurden hier für die Deskriptor-Gitter BoF-Sequenzen verwendet. Das sind jeweils 5 Pixel in horizontaler und vertikaler Richtung für den GW Datensatz und 2 Pixel für den JB-V Datensatz. Um die Vergleichbarkeit zwischen den BoF-Sequenzen und den D-Sequenzen zu gewährleisten, wurden die gleichen Schrittweiten für das gleitende Fenster verwendet. Weiterhin wurden in den Basisexperimenten die in [SF15] ermittelten Größen für das visuelle Vokabular beider BoF-Sequenzen verwendet. Für den GW Datensatz sind das 4096 Einträge und für den JB-V Datensatz 1024 Einträge.

Bag-of-Features Sequenzen

Die Ergebnisse der Basisexperimente zeigen, dass sich Kosinus-, L1- und die BC-Distanz besser als die L2-Distanz für den Vergleich der Bag-of-Features Histogramme (BoF-Histogramme) im DTW eignen. Die mAP-Werte mit der L2-Distanz sind für beide

	GW		JB-V	
Distanz	SIFT-BoF	LGH-BoF	SIFT-BoF	LGH-BoF
L2	14,4	10,2	33,2	24,5
Kosinus	**51,9**	**48,8**	69,6	68,6
L1	49,6	46,4	71,9	71,3
BC	51,2	48,1	**72,5**	**72,3**

Tabelle 5.2.1: Mean Average Precisions (in %) der Basisexperimente für die BoF-Sequenzen.

Datensätze bedeutend schlechter (siehe Tabelle 5.2.1). Vergleicht man die Ergebnisse weiter, so stellt man fest, dass mit der BC-Distanz in allen Fällen bessere mAP-Werte erreicht werden als mit der L1-Distanz. Bei der Anwendung der BC-Distanz sei gesagt, dass die Normalisierung der BoF-Histogramme (siehe Kapitel 4.2) nicht durchgeführt wurde. Entsprechend unterscheiden sich L1- und BC-Distanz hier insofern, dass bei der BC-Distanz die Höhenunterschiede zwischen den Wortabbildern berücksichtigt werden. Die Höhenunterschiede beeinflussen das Word Spotting Ergebnis hier offensichtlich positiv.
Für den JB-V Datensatz hat die BC-Distanz insgesamt am besten abgeschnitten, für den GW Datensatz die Kosinus-Distanz (siehe Tabelle 5.2.1). Eine weitere Erkenntnis der Basisexperimente ist, dass der SIFT-Deskriptor im Vergleich zum LGH-Deskriptor bessere Ergebnisse liefert. Beide Deskriptoren wurden nach [SF15] auf eine Größe von 40×40 Pixel beim GW Datensatz und 24×24 Pixel beim JB-V Datensatz eingestellt. Die Deskriptoren waren jeweils in 4×4 Zellen eingeteilt und die Anzahl der Hauptorientierungen der Gradientenhistogramme war auf 8 festgelegt. Die besseren Ergebnisse des SIFT-Deskriptors lassen sich auf die in Kapitel 3.1.1 genannten Unterschiede zurückführen. Vor allem die Gaußgewichtung der Bildgradienten im Randbereich des SIFT-Deskriptors wird hier als Grund für die besseren Resultate vermutet. Diese macht den SIFT-Deskriptor gegenüber dem LGH-Deskriptor weniger spezifisch, für den jeweiligen Bildbereich. Dadurch ist der SIFT-Deskriptor vermutlich weniger anfällig gegenüber Kontrastvariationen des Hintergrundes und bietet daher eine bessere Beschreibung der Schriftverläufe.

Deskriptor Sequenzen

Für den späteren direkten Vergleich zu den BoF-Sequenzen wurde die Fensterbreite in den Basisexperimenten für den GW Datensatz auf 40 Pixel und für den JB-V Datensatz

	GW			JB-V		
Distanz	LGH-Full	LGH-Niblack	VIN-Niblack	LGH-Full	LGH-Niblack	VIN-Niblack
L2	36,3	43,2	32,2	54,9	66,2	64,1
Kosinus	37,6	46,0	33,5	56,8	71,8	**66,8**
L1	**37,9**	**48,0**	**33,7**	59,7	**72,6**	66,3

Tabelle 5.2.2: Mean Average Precisions (in %) der Basisexperimente für die LGH-Full, LGH-Niblack und VIN-Niblack Varianten der D-Sequenz.

auf 24 Pixel eingestellt. Auf diese Weise ist für alle verschiedenen sequentiellen Wortrepräsentationen (siehe Kapitel 4.4) eines Wortabbildes sichergestellt, dass die jeweiligen Merkmalsvektoren der Sequenzen auf den gleichen lokalen Bildausschnitten basieren. Der LGH-Deskriptor wird hier, wie in [RP08] und im Methodikkapitel 4.1 beschrieben, in der einfachen Variante (LGH-Full) und in der angepassten Variante (LGH-Niblack) mit 4×4 Zellen und 8 Hauptorientierungen getestet. Für den VIN-Deskriptor wird die Zellenaufteilung hier ebenfalls auf 4×4 eingestellt. Erste informelle Experimente zum Parameter α für den Niblack-Schwellwert der Binarisierung (siehe Kapitel 2.1 und 4.1) zeigten gute Resultate für $\alpha = -0,2$ auf beiden Datensätzen. Dieser wird daher für die weiteren Experimente beibehalten.
Die Ergebnisse der Basisexperimente zu den D-Sequenzen (siehe Tabelle 5.2.2) bestätigen die Erkenntnis aus [RP08]. Die Anpassung des LGH-Deskriptors auf den Schriftbereich führt auch hier zu deutlich besseren Resultaten auf beiden Datensätzen (siehe Tabelle 5.2.2). Im weiteren Verlauf der Arbeit wird die LGH-Full Repräsentation daher nicht weiter betrachtet. Zu dem Erfolg dieser Anpassung muss hier gesagt werden, dass sie und damit auch die Word Spotting Resultate stark von der Qualität der Binarisierung abhängen. Frühe Experimente auf dem JB-V Datensatz mit einer hier ungeeigneten Binarisierung mit dem Otsu-Schwellwert [Ots79] führten zu weniger guten Resultaten (siehe Tabelle A.0.4).
Weiterhin zeigen die Basisexperimente, dass sich die LGH-Niblack Variante der D-Sequenz hier besser als die VIN-Niblack Variante zur Beschreibung der Schrift, auf beiden Datensätzen, eignet (siehe Tabelle 5.2.2). Allerdings sei gesagt, dass die VIN-Niblack Repräsentation deutlich geringer in ihrer Dimensionalität ist, als die LGH-Niblack Repräsentation (vgl. Kapitel 3.1.2). Zusätzliche Experimente zur LGH-Niblack Variante mit nur einer Hauptorientierung führten in der Regel zu schlechteren mAP-Werten als die vergleichbare VIN-Niblack Variante (siehe Tabelle A.0.7). Der Einsatz der VIN-Niblack Variante der D-Sequenz kann also für manche Anwendungen aus

Laufzeitgründen gerechtfertigt sein. Außerdem lässt sich festhalten, dass die Orientierungsinformationen der Gradientenhistogramme wichtig für den Erfolg der LGH-Niblack Repräsentation sind.
In den Basisexperimenten wurde die L1-Distanz als bestes Distanzmaß für den LGH-Deskriptor auf beiden Datensätzen ermittelt (siehe Tabelle 5.2.2). Auch für die VIN-Niblack Variante ergab sich das beste Ergebnis auf dem GW Datensatz mit der L1-Distanz. Für den JB-V Datensatz wurde mit der Kosinus-Distanz in der Konfiguration der Basisexperimente eine minimal bessere mAP erzielt (siehe Tabelle 5.2.2). Diese Abweichung ist allerdings klein und wird daher als nicht signifikant angesehen. Auch in weiteren Experimenten konnte kein deutlicher Vorteil der Kosinus-Distanz erkannt werden. Für die spätere Parameterauswertung wird daher auch für die VIN-Niblack Variante die L1-Distanz verwendet.

5.2.2 *Parameterauswertung*

Bei der Parameterauswertung wird hier versucht, durch gezielte Experimente die mAP-Werte, welche in den Basisexperimenten erzielt wurden, weiter zu steigern. Die Experimente beschränken sich auf die wichtigsten Parameter der jeweiligen Methodik, mit dem Fokus auf den Parametern der Merkmale. Neben der Verbesserung der mAP-Werte sollen die Experimente tieferen Aufschluss über die Einflüsse der jeweiligen Parameter geben, um einen differenzierteren Vergleich zwischen BoF-Sequenz und D-Sequenz in Kapitel 5.3 zu ermöglichen.

Bag-of-Features Sequenzen

Bei einem großen visuellen Vokabular sind die einzelnen visuellen Worte spezifischer als bei kleineren visuellen Vokabularen (vgl. Kapitel 4.2). Für die in [ARTL15] vorgeschlagene holistische SP-BoF Repräsentation (vgl. Kapitel 3.2) wurde der Vorteil eines großen visuellen Vokabulars bestätigt.
Durch die sequentielle Repräsentation und die ausgleichende Wirkung des DTW (vgl. Kapitel 2.4) lässt sich vermuten, dass die einzelnen BoF-Histogramme der BoF-Sequenzen hier weniger spezifisch sein können und trotzdem gute mAP-Werte erzielt werden. Das visuelle Vokabular kann also kleiner sein als in den Basisexperimenten ausgewählt. Experimente bekräftigen diese Hypothese (siehe Tabelle 5.2.3). So wurden auf dem JB-V Datensatz für die SIFT-BoF und die LGH-BoF Variante die besten Ergebnisse mit 512 visuellen Worten erzielt. Auf dem GW Datensatz wurden die besten Ergebnisse bei 2048 bzw. 1024 visuellen Worten für die SIFT- bzw. LGH-BoF

	GW (Kosinus)		JB-V (BC)	
Vokabulargröße	SIFT-BoF	LGH-BoF	SIFT-BoF	LGH-BoF
4096	51,9	48,8	-	-
2048	**53,9**	50,4	70,2	-
1024	53,8	**50,6**	72,5	72,3
512	52,8	48,5	**74,1**	**73,1**
256	51,3	47,0	72,9	71,5

Tabelle 5.2.3: Mean Average Precisions (in %) für verschiedene Größen des visuellen Vokabulars für SIFT-BoF und LGH-BoF Sequenzen.

Variante erreicht (siehe Tabelle 5.2.3). Das kleinere visuelle Vokabular verringert zwar die Dimensionalität der BoF-Sequenzen, dennoch bleibt die Dimensionalität groß.

Deskriptor Sequenzen

Nach den Ergebnissen der Basisexperimente werden hier noch die LGH-Niblack Sequenz und die VIN-Niblack Sequenz betrachtet.
Sowohl der VIN-Deskriptor als auch der LGH-Deskriptor wurden mit einer Einteilung in 4×4 Zellen vorgeschlagen (vgl. [AJoo], [RPo8]). Es lässt sich vermuten, dass eine andere Zelleneinteilung hier zu besseren mAP-Werten führen kann. Es wurden daher Experimente zur Zelleneinteilung der Deskriptoren durchgeführt. Außerdem wurden Experimente zur Fensterbreite durchgeführt, da diese erheblichen Einfluss auf die Anpassung des Deskriptors auf den Schriftbereich und die Breite der Zellen der Deskriptoren haben.

LGH-Niblack Sequenz

Experimente zu verschiedenen Zelleneinteilungen haben gezeigt, dass sich Verbesserungen der mAP durch eine andere Zelleneinteilung für die LGH-Niblack Sequenz ergeben können. So wurde auf dem JB-V Datensatz bei einer Fensterbreite von 24 Pixeln die beste mAP bei einer Zelleneinteilung von 5×6 Zellen erreicht. Auf dem GW Datensatz wurde bei einer Fensterbreite von 40 Pixeln die beste mAP bei einer Zelleneinteilung von 3×5 gefunden (siehe Tabelle 5.2.4). Eine feinere Zelleneinteilung des LGH-Deskriptors in Schriftrichtung kann also bei ausreichender Fensterbreite zu besseren Ergebnissen führen. Über die vertikale Einteilung lässt sich nur schwer

GW (Fensterbreite 40 Pixel)				JB-V (Fensterbreite 24 Pixel)			
	Spalten				Spalten		
Zeilen	4	5	8	Zeilen	4	6	8
2	47,2	47,5	47,2	4	72,6	73,8	73,9
3	48,2	**48,5**	48,3	5	73,0	**74,0**	73,9
4	48,0	48,1	48,0	6	72,6	73,2	73,1

Tabelle 5.2.4: Mean Average Precisions (in %) für verschiedene Zelleneinteilungen des LGH-Deskriptors.

eine Aussage treffen, da die Höhe der Zellen maßgeblich von der Anpassung des Deskriptors auf den Schriftbereich abhängt und somit für die einzelnen Deskriptoren verschieden ist.
Für die Fensterbreite wird vermutet, dass eine geringere Fensterbreite zu einer besseren Anpassung des Deskriptors auf den Schriftbereich und somit zu besseren mAP führt. Die Ergebnisse der Experimente bestätigen diese Vermutung. So wurden bei einer Zelleneinteilung in 4×4 Zellen die besten mAP-Werte bei Fensterbreiten von 16 Pixeln auf dem JB-V Datensatz und 24 Pixeln auf dem GW Datensatz ermittelt (siehe Tabelle 5.2.5). Durch die kleinere Fensterbreite verliert sich zudem der positive Effekt einer feineren Zelleneinteilung des Deskriptors. Bei erneuten Experimenten zur Zelleneinteilung bei diesen Fensterbreiten ergaben sich keine nennenswerten Verbesserungen der mAP.
Während der Parameterauswertung wurden zusätzliche Experimente zur Anzahl der Hauptorientierungen durchgeführt (siehe Anhang A.0.6). Diese haben gezeigt, dass sich die mAP auf dem GW Datensatz durch eine Erhöhung dieser Anzahl steigern lässt. Diese Ergebnisse decken sich mit den Ergebnissen aus [TT09]. Dort wurde ein ähnlicher Deskriptor in einem auf Zeilensegmentierung basierten Word Spotting Szenario vorgeschlagen.
Für den JB-V Datensatz blieb eine Steigerung durch Erhöhung der Anzahl der Hauptorientierungen aus (siehe Anhang A.0.6). Das unterschiedliche Verhalten lässt sich dadurch erklären, dass die Schriftvariation in dem GW Datensatz relativ gering ist. Bei den JB Datensätzen sind die Variationen größer, der Deskriptor sollte also etwas stärker abstrahieren.

GW		JB-V	
Fensterbreite	mAP	Fensterbreite	mAP
20	50,3	12	73,7
24	**50,5**	16	**75,3**
28	50,1	20	74,1
32	49,6	24	72,6

Tabelle 5.2.5: Mean Average Precisions (in %) für verschiedene Fensterbreiten der LGH-Niblack Sequenz, bei einer Zelleneinteilung in 4×4 Zellen.

VIN-Niblack Sequenz

Auch für die VIN-Niblack Sequenz wurden Experimente zur Zelleneinteilung und Fensterbreite durchgeführt. Diese Experimente zeigen, dass eine feinere Zelleneinteilung ebenfalls zu besseren Ergebnissen führen kann. Für den GW Datensatz ergab sich die beste mAP von **35,9** bei einer Fensterbreite von 40 Pixeln und einer Unterteilung des VIN-Deskriptors in 4×20 Zellen (siehe Anhang A.0.8). Das führt zu einem 120-dimensionalen Vektor. Das Laufzeitargument für den VIN-Deskriptor ist damit hinfällig. Auch die mAP-Werte rechtfertigen keine weiteren Experimente auf dem GW Datensatz. Die Steigerung der mAP durch Erhöhung der Spaltenauflösung lässt sich damit erklären, dass es sich bei dem GW Datensatz um ein Einzelschreiber Szenario handelt. Die spezifischere Beschreibung der Schrift hat hier offenbar positiven Einfluss. Für den JB-V Datensatz ist die beste ermittelte Konfiguration eine Fensterbreite von 24 Pixeln zusammen mit einer Deskriptoraufteilung in 5×8 Zellen. Hierfür ergab sich eine mAP von **69,6** (siehe Anhang A.0.8). Eine feinere Zelleneinteilung führte zu schlechteren Resultaten. Das lässt sich damit erklären, dass die Schriftvariation beim JB-V Datensatz höher ist. Durch Veränderungen der Fensterbreite ließ sich die mAP nicht weiter steigern (siehe Anhang A.0.5).
Im Vergleich zum LGH-Deskriptor zeigt der VIN-Deskriptor insgesamt ein schlechteres Verhalten auf den betrachteten Datensätzen. Das kann daran liegen, dass der VIN-Deskriptor stärker von der Binarisierung abhängt und diese nicht optimal ist. Oder auch daran, dass Deskriptoren auf Basis von Gradientenhistogrammen hier die Eigenschaften von Handschriften besser repräsentieren. Darauf deutet auch die aktuelle Literatur zum Word Spotting hin. Repräsentationen auf Basis von Gradientenhistogrammen sind hier häufig vertreten (z.B. [RRLF14], [RP09], [AGFV14]).

5.3 BAG-OF-FEATURES SEQUENZ VS. DESKRIPTOR SEQUENZ

Nachdem die BoF-Sequenz und D-Sequenz mit ihren jeweiligen Merkmalen unabhängig voneinander betrachtet wurden, folgt ein Vergleich der beiden Repräsentationen miteinander. Wie in Abbildung 3.2.1 angedeutet, werden dabei die LGH-Niblack und die LGH-BoF betrachtet.
Mit der LGH-BoF Repräsentation wurden auf dem **GW** Datensatz bessere Ergebnisse als mit der LGH-Niblack Repräsentation erzielt, sowohl in den Basisexperimenten als auch nach der Parameterauswertung. Nach den Basisexperimenten ergab sich eine mAP von 48,8 für die LGH-BoF Repräsentation (vgl. Tabelle 5.2.1) und eine mAP von 48,0 für die LGH-Niblack Repräsentation (vgl. Tabelle 5.2.2) als beste Werte. Bei der Parameterauswertung wurde die höchste mAP für die LGH-BoF Repräsentation bei 50,6 erzielt (vgl. Tabelle 5.2.3). Für die LGH-Niblack Repräsentation lag diese bei 50,5 (vgl. Tabelle 5.2.5). Obwohl die Parameterauswertung für die LGH-Niblack Repräsentation ausführlicher ausfiel, wurde das Ergebnis der LGH-BoF Repräsentation nicht übertroffen.
Ein Grund dafür sind vermutlich die segmentierten Wortabbilder des GW Datensatzes. In diese Wortabbilder ragen häufig Ober- oder Unterlängen umliegender Worte herein. Das verhindert die korrekte Anpassung des Deskriptors auf den Schriftbereich des Wortes. In den BoF-Sequenzen stören solche Artefakte weniger wegen der Anordnung der Deskriptoren im dichten Gitter. Es werden nicht alle charakteristischen Merkmale durch diese Artefakte beeinflusst. Die betroffenen BoF-Histogramme sind also weiterhin hinreichend charakteristisch für das Wort. Bei der D-Sequenz werden durch die fehlerhafte Anpassung alle Einträge im jeweiligen Merkmalsvektor verfälscht. Abbildung 5.3.1 zeigt ein Beispielergebnis für das Wort "Winchester" und soll ein Indiz für die vorherige Hypothese sein. Die Ergebnisse stammen jeweils aus der besten Parameterkonfiguration der LGH-Niblack Repräsentation und der LGH-BoF Repräsentation.
Auf dem **JB-V** Datensatz ergaben sich mit der LGH-Niblack Repräsentation bessere mAP-Werte, sowohl nach den Basisexperimenten als auch nach der Parameterauswertung. Nach den Basisexperimenten ergab sich eine mAP von 72,5 für die LGH-BoF Repräsentation (vgl. Tabelle 5.2.1) und eine mAP von 72,6 für die LGH-Niblack Repräsentation (vgl. Tabelle 5.2.2) als beste mAP-Werte. Bei der Parameterauswertung wurde die höchste mAP für die LGH-BoF Repräsentation bei 73,1 erzielt (vgl. Tabelle 5.2.3). Für die LGH-Niblack Repräsentation lag diese bei 75,3 (vgl. Tabelle 5.2.5). Als Grund für die besseren Ergebnisse wird hier die Anpassung der Deskriptoren auf den Schriftbereich gesehen. Die segmentierten Wortabbilder enthalten auf dem JB-V Datensatz deutlich weniger Ober- bzw. Unterlängen anderer Worte, die Anpassung

Abbildung 5.3.1: Qualitative Ergebnisse eines Anfragebildes zum Wort "Winchester" auf dem GW Datensatz.

gelingt hier vermutlich besser.
Eine klare Aussage zur Leitfrage (siehe Kapitel 1.2), ob sich die BoF-Sequenz besser für Word Spotting in historischen Dokumenten eignet als die D-Sequenz, lässt sich an dieser Stelle nicht treffen. Allerdings zeigen die Ergebnisse auf dem GW Datensatz, dass die BoF-Sequenz robuster im Bezug auf oben genannte Artefakte innerhalb der segmentierten Wortabbildern sein kann.

5.4 VERGLEICH ZUM STATE-OF-THE-ART

Zunächst wird auf die Auswertungsergebnisse zum GW Datensatz eingegangen. Dabei werden einige Eigenschaften des DTW-Algorithmus der vorliegenden Arbeit untersucht. Anschließend werden die Auswertungsergebnisse auf dem JB-V und JB-E Datensatz betrachtet. Der Vergleich bezieht sich auf die Qualität der Ergebnisse. Laufzeitaspekte werden hier nicht betrachtet.

5.4.1 *George Washington*

Die beste bisher vorgestellte mAP dieser Arbeit liegt bei **53,9** und ergab sich mit der SIFT-BoF Repräsentation (vgl. Tabelle 5.2.3). In [SF15] wurde eine mAP von **60,59** mit einer SP-BoF Repräsentation, wie in Kapitel 3.2 beschrieben (ohne Dimensionsreduktion), erreicht. Die zuvor angegeben mAP aus [SF15] entstand durch eine Ähnlichkeitsbewertung durch der Kosinus-Distanz. Da die Parameterwerte in dieser Arbeit durch die Parameterwerte aus [SF15] beeinflusst wurden (siehe Kapitel 5.2.1),

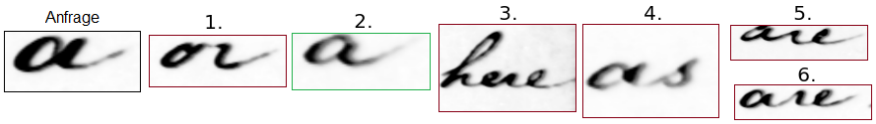

Abbildung 5.4.1: Qualitatives Ergebnis der Anfrage mit einem Wortabbild zum Wort ”a” auf dem GW Datensatz mit der LGH-Niblack Repräsentation.

lassen sich die mAP-Werte von 53,9 hier und 60,59 in [SF15] am ehesten vergleichen. Als Grund für die geringen mAP-Werte dieser Arbeit auf dem GW Datensatz wird der DTW-Algorithmus der vorliegenden Arbeit gesehen. In den Experimenten fiel auf, dass die Ergebnisse für kurze Anfragewörter deutlich schlechter waren als die für längere Anfragewörter (siehe Tabellen A.0.1, A.0.2).
Abbildung 5.4.1 zeigt ein qualitatives Ergebnis für das Anfragewort ”a” auf dem GW Datensatz. Das Ergebnis wurde mit der LGH-Niblack Repräsentation erzielt, die Resultate zur LGH-BoF und die SIFT-BoF waren ähnlich. Es fällt auf, dass die Anfrageergebnisse dem Anfragebild in weiten Teilen ähneln. Im Beispiel ist das Anfragewort in fast allen der vordersten Worte der Antwortliste enthalten. Für kurze Wörter ist dies häufig der Fall. Da das DTW versucht, eine optimale Zuordnung zwischen den Sequenzen zu finden (siehe Kapitel 2.4), haben ähnliche Teilstücke vermutlich einen positiveren Einfluss auf den Ähnlichkeitswert als unähnliche Teilstücke.
Experimente mit einer geringeren Schrittweite des gleitenden Fensters haben gezeigt, dass sich die Ergebnisse kurzer Wörter verbessern lassen (siehe Tabelle A.0.2). Durch die geringere Schrittweite wird der Informationsgehalt der sequentiellen Repräsentation erhöht, was das DTW begünstigt.
Experimente zum Anpassungsfenster des DTW auf dem GW Datensatz haben weiterhin gezeigt, dass das Anpassungsfenster negativen Einfluss auf die Resultate kurzer Anfragewörter hat (siehe Tabelle A.0.1). Die Menge der Warping Funktionen wird durch das Anpassungsfenster vermutlich so beschränkt, dass die unähnliche Mitte, wie beispielsweise für die beiden Wortabbilder zu ”are” in Abbildung 5.4.1, kaum ins Gewicht fällt.
Für Wörter mit mehr als sechs Buchstaben waren die Auswirkungen dieser Parameter in den Experimenten weniger stark (siehe Tabellen A.0.1 und A.0.2).
Es bleibt festzuhalten, dass DTW hier Schwächen beim Word Spotting mit kurzen Anfragenwörtern aufweist. Kurze Anfragewörter sind jedoch nicht unbedingt für jede Anwendung relevant, so lässt sich das Evaluierungsprotokoll für den GW Datensatz in Frage stellen.

5.4.2 *Jeremy Bentham*

Methode	mAP	Parameter
Methode der PRG	42,44	SIFT basierte SP-BoF, für weitere Details siehe [PTV15]
Methode der CVC	30,00	IHOG basierte SP-BoF, für weitere Details siehe [PTV15]
SIFT-BoF	**44,31**	24 Pixel Deskriptorgröße, 4×4 Zellen, Deskriptorabstand 2 Pixel
LGH-BoF	41,47	24 Pixel Deskriptorgröße, 4×4 Zellen, Deskriptorabstand 2 Pixel
LGH-Nibalck	**43,18**	16 Pixel Deskriptorbreite, 2 Pixel Schrittweite, 4×4 Zellen
VIN-Nibalck	37,20	24 Pixel Fensterbreite, 2 Pixel Schrittweite, 5×8 Zellen

Tabelle 5.4.1: Vergleich der Mean Average Precisions (in %) des Wettbewerbes zum "Keyword Spotting for Handwritten Documents" [PTV15].

Anders als das Evaluierungsprotokoll auf dem GW Datensatz enthalten die Anfragewörter sowohl für den JB-V als auch für den JB-E Datensatz nur Wortabbilder von Worten mit sechs oder mehr Buchstaben (siehe Kapitel 5.1). Auf beiden untersuchten Ausschnitten der "Bentham Collection" konnten die Ergebnisse anderer moderner Methoden übertroffen werden. So wurde die beste mAP von **72,63** aus [SF15] für den JB-V Datensatz, sowohl mit den BoF-Sequenzen als auch mit der LGH-Niblack Repräsentation, übertroffen (vgl. Tabellen 5.2.3, 5.2.5). Die beste mAP der VIN-Niblack Repräsentation liegt mit **69,6** (siehe Anhang A.0.8) knapp darunter.
Die Auswertung auf dem JB-E Datensatz zeigt ebenfalls, dass die hier beschriebene Methodik bezüglich ihrer Ergebnisqualität auf dem Niveau anderer moderner Methoden liegt. Mit der SIFT-BoF und der LGH-Niblack Repräsentation in Kombination mit dem eingesetzten DTW-Algorithmus konnte das beste Ergebnis des Wettbewerbs zum Word Spotting auf der ICDAR2015 [PTV15] übertroffen werden (vgl. Tabelle 5.4.1). Auch die VIN-Niblack Repräsentation zeigt hier gute Resultate, so konnte die Methode der CVC Gruppe (siehe Kapitel 3.2) übertroffen werden (vgl. Tabelle 5.4.1).
Als Vorteil der Methodik in der vorliegenden Arbeit wird vor allem die sequentielle Repräsentation gesehen. Durch das gleitende Fenster und die dichte Schrittweite ist die räumliche Information sehr detailliert erfasst. Gleichzeitig repräsentiert die Sequenz die Schrift hier intuitiv natürlicher als die Spatial Pyramid, welche in [SF15] und den Methoden der PRG (siehe Kapitel 3.2) und des CVC verwendet werden.

6 FAZIT

In dieser Arbeit wurden zwei verschiedene sequentielle Wortrepräsentationen, die BoF-Sequenz und die D-Sequenz, in Kombination mit verschiedenen Histogramm-basierten Merkmalen zum Word Spotting auf historischen Dokumenten untersucht. Die Merkmalsextraktion beider Sequenzen folgt dabei unterschiedlichen Ansätzen. Bei der D-Sequenz werden die Merkmalsvektoren direkt als lokale Bilddeskriptoren extrahiert (siehe Kapitel 4.1). Bei der BoF-Sequenz wird vorher ein Abstraktionsschritt mit Informationen über einen Teil der untersuchten Daten durchgeführt. Bei diesem Abstraktionsschritt werden lokale Bilddeskriptoren zu charakteristischen Merkmalen zusammengefasst und anschließend in lokalen BoF-Histogrammen zur Repräsentation verwendet (siehe Kapitel 4.2). Beide Repräsentationen wurden bereits in anderen Arbeiten zum Word Spotting auf anderen Datensätzen oder in anderen Szenarien und Systemen eingesetzt (siehe Kapitel 3).
Als Grundlage für die D-Sequenz und andere Bag-of-Features Ansätze wurden in anderen Arbeiten verschiedene lokale Bilddeskriptoren zum Word Spotting ausgewertet (siehe Kapitel 3 und 4). Eine Auswertung der D-Sequenz auf Datensätzen historischer Dokumente und ein direkter Vergleich zwischen der D-Sequenz und der BoF-Sequenz ist durch diese Arbeit gegeben.
Einige Erkenntnisse aus [RP08] werden hier durch die Auswertung der D-Sequenz auf dem GW Datensatz und den JB Datensätzen bestätigt. Zunächst konnte der positive Effekt der Anpassung des LGH-Deskriptors auf den Schriftbereich auch für diese Datensätze bestätigt werden (siehe Kapitel 5.2.1). Weiterhin wurden hier, wie in [RP08], mit dem LGH-Deskriptor bessere Ergebnisse im Vergleich zum VIN-Deskriptor erzielt. Gradientenhistogramm-basierte Merkmale scheinen besser zum Word Spotting auf historischen Dokumenten geeignet zu sein (siehe Kapitel 5.2.2).
Für die BoF-Sequenz wurde der LGH-Deskriptor im Vergleich zum SIFT-Deskriptor ausgewertet. Für den SIFT-Deskriptor ergaben sich dabei die besseren Ergebnisse. Als Grund dafür wird hier die Abschwächung der Gradienten im Randbereich des SIFT-Deskriptors vermutet (vgl. Kapitel 5.2.1). Dadurch ist der SIFT-Deskriptor weniger anfällig gegen Kontrastvariationen im Hintergrund der Dokumentenabbilder und bietet daher eine allgemeinere Beschreibung für die Schrift.
Weiterhin wurden in dieser Arbeit einige Schwächen des DTW im Bezug auf kurze Wörter beim GW Datensatz aufgedeckt (siehe Kapitel 5.4.1). So eignet sich das DTW

hier nur bedingt für die Suche nach kleinen Worten. Für den JB Datensatz und das Evaluierungsprotokoll konnten die LGH-Niblack und die SIFT-BoF Repräsentation die Ergebnisse anderer moderner Ansätze zum Word Spotting übertreffen (siehe Kapitel 5.4.2). Als Grund dafür wird hier die natürliche Beschreibung der Schrift durch die sequentiellen Wortrepräsentationen in Kombination mit dem DTW gesehen.
An dieser Stelle sei gesagt, dass die durchschnittliche Dimensionalität der LGH-Niblack und BoF-Sequenzen relativ groß ist. Zusammen mit der hohen Laufzeit des DTW-Algorithmus ist der Einsatz auf Datensätzen mit vielen Dokumenten in der hier vorgestellten Form vermutlich nicht praktikabel. Wie auch in anderen Arbeiten zum segmentierungsbasierten Word Spotting könnten allerdings Pruning Strategien angewendet werden, um die Laufzeit zu verringern. Zusätzlich wäre der Einsatz effizienterer Lösungen für das DTW denkbar (z.B. [SC07]). Um die Dimensionalität der BoF-Sequenzen zu verringern wäre zudem der Einsatz von Verfahren zur Dimensionsreduktion der einzelnen BoF-Histogramme denkbar, ähnlich wie in [RATL11] und [SF15] für SP-BoF Repräsentationen vorgeschlagen.
Der Vergleich zwischen D-Sequenz und BoF-Sequenz ergab, dass die Ergebnisse auf den untersuchten Datensätzen auf einem ähnlichen Niveau liegen. Vorausgesetzt man verwendet für beide den LGH-Deskriptor und ein geeignetes Distanzmaß für das DTW. Je nach Datensatz zeigen sich leichte Vorteile für die LGH-BoF bzw. die LGH-Niblack Repräsentation. So erreicht die LGH-Niblack Repräsentation bessere Ergebnisse auf den JB Datensätzen. Als Grund dafür wird die Anpassung des LGH-Deskriptors auf den Schriftbereich gesehen (vgl. Kapitel 5.2.1). Diese gelingt auf den JB Datensätzen wegen der *klaren Segmentierung*. Mit klar ist hier gemeint, dass die segmentierten Wortabbilder keine oder nur selten störende Artefakte wie Ober- oder Unterlängen anderer Wörter enthalten. Gelingt die Anpassung auf den Schriftbereich weniger gut, wie auf dem GW Datensatz, so zeigen die LGH-BoF und die SIFT-BoF Repräsentation ein robusteres Verhalten. Für die Antwort auf die anfangs gestellte Leitfrage, ob die D-Sequenz oder die BoF-Sequenz besser zum Word Spotting in historischen Dokumenten geeignet ist (siehe Kapitel 1.2), muss hier also differenziert werden:
Die Wahl einer passenden Repräsentation hängt stark von der Datengrundlage ab. Erlauben die Dokumente eine klare Segmentierung und eine gute Binarisierung, im dem Sinne, dass die Anpassung des LGH-Deskriptors auf den Schriftbereich gut funktioniert, so kann die LGH-Niblack Variante der D-Sequenz eine gute Wahl zum Word Spotting sein. Enthalten die segmentierten Wortabbilder allerdings häufiger störende Artefakte, wie zuvor beschrieben, so sind die BoF-Sequenzen, insbesondere die SIFT-BoF Sequenz, die bessere Wahl. Durch die Anordnung der Deskriptoren im dichten Gitter werden die BoF-Sequenzen als weniger anfällig gegen solche Störungen gesehen (vgl. Kapitel 5.3). Insbesondere für Word Spotting in historischen Dokumen-

ten können die SIFT-BoF-Sequenzen also interessant sein, da die Qualität solcher Dokumente durch Alterungsprozesse und andere äußere Einflüsse beeinträchtigt sein kann. Selbst dann, wenn die Qualität der Dokumente so stark beeinträchtigt ist, dass gar keine Segmentierung möglich ist, lassen sich die BoF-Sequenzen auch in einem segmentierungsfreien Szenario anwenden [RRLF14], [FRG14].

A ANHANG

PARAMETERLISTEN ZUR METHODIK

Hier wird eine Übersicht über alle einstellbaren Parameter der Methodik gegeben.

Bag-of-Features Sequenz

Dichtes Deskriptor Gitter

- horizontaler Abstand zwischen Deskriptoren in Pixeln
- vertikaler Abstand zwischen Deskriptoren in Pixeln
- Höhe der Deskriptoren in Pixeln
- Breite der Deskriptoren in Pixeln
- Anzahl der Deskriptor Zeilen
- Anzahl der Deskriptor Spalten

Gleitendes Fenster

- Fensterschrittweite auf dem Merkmalsgitter
- Fensterbreite auf dem Merkmalsgitter

Bag-of-Features

- Größe des visuellen Vokabular

Dynamic-Time-Warping

- Distanzmaß [L2, Kosinus, L1, BrayCurtis]
- Breite des Anpassungsfenster

Deskriptor Sequenz

Deskriptoren

- Anzahl der Deskriptor Zeilen
- Anzahl der Deskriptor Spalten
- Binarisierungsparameter α für Anpassung und VIN-Deskriptor
- Anzahl der Hauptorientierungen der Gradientenhistogramme

Gleitendes Fenster

- Fensterschrittweite in Pixeln
- Fensterbreite in Pixeln

Dynamic-Time-Warping

- Distanzmaß [L2, Kosinus, L1]
- Anpassungsfensterbreite

WEITERE AUSWERTUNGSERGEBNISSE

Hier sind weitere Auswertungsergebnisse aufgeführt, die entweder zu detailliert für das Evaluierungskapitel sind oder bei welchen es sich um Stichprobenexperimente handelt, um Einflüsse anderer Parameter zu prüfen.

Einfluss des Anpassungsfensters

Diese Experimente wurden mit der SIFT-BoF Repräsentation durchgeführt.
Folgende Parameter waren für den GW Datensatz eingestellt: Visuelles Vokabular der Größe 4096, Deskriptorgröße 40 Pixel, Fensterschrittweite 5 Pixel in x- und y-Richtung, Fensterbreite 1 Deskriptorbreite bzw. 40 Pixel, BC-Distanz für das DTW.
Folgende Parameter waren für den JB-V Datensatz eingestellt: Visuelles Vokabular der Größe 1024, Deskriptorgröße 24 Pixel, Fensterschrittweite 2 Pixel in x- und y-Richtung, Fensterbreite 1 Deskriptorbreite bzw. 24 Pixel, BC-Distanz für das DTW.

Anpassungsfenster	GW	JB-V
Nein	57,1 / 54,1 / 66,8	71,7
Ja (10%)	51,2 / 46,5 / 66,8	72,7

Tabelle A.0.1: Mean Average Precisions der Experimente zum Einfluss des Anpassungsfensters. Für den GW Datensatz sind jeweils die gesamte mAP, die mAP der Anfragen kurzer Worte (1-5 Zeichen) und die mAP der Anfragen langer Worte (6 oder mehr Zeichen), in dieser Reihenfolge angegeben.

Einfluss der Fensterschrittweite

Diese Experimente wurden mit der LGH-Niblack Repräsentation durchgeführt.
Folgende Parameter waren für den GW Datensatz eingestellt: $\alpha = -0.2$, Fensterbreite 24 Pixel, 4×4 Zellen, 8 Hauptorientierungen, L1-Distanz für das DTW, Anpassungsfenster 10%.
Folgende Parameter waren für den JB-V Datensatz eingestellt: $\alpha = -0.2$, Fensterbreite 24 Pixel, 4×4 Zellen, L1-Distanz für das DTW, Anpassungsfenster 10%.

Fensterschrittweite	GW	JB-V
5 Pixel	50,5 / 46,7 / 63,0	69,1
2 Pixel	57,9 / 55,4 / 66,2	72,6

Tabelle A.0.2: Mean Average Precisions der Experimente zum Einfluss der Fensterschrittweite. Für den GW Datensatz sind jeweils die gesamte mAP, die mAP der Anfragen kurzer Worte (1-5 Zeichen) und die mAP der Anfragen langer Worte (6 oder mehr Zeichen), in dieser Reihenfolge, angegeben.

Einfluss der DTW Normalisierung

Gemeint ist die Normalisierung durch die Länge des Warping Pfad. Diese Experimente wurden mit der SIFT-BoF Repräsentation durchgeführt.
Folgende Parameter waren für den GW Datensatz eingestellt: Visuelles Vokabular der Größe 4096, Deskriptorgröße 40×40 Pixel, Fensterschrittweite 5 Pixel in x- und y-Richtung, Fensterbreite 1 Deskriptorbreite bzw. 40 Pixel, BC-Distanz für das DTW und kein Anpassungsfenster.

Folgende Parameter waren für den JB-V Datensatz eingestellt: Visuelles Vokabular der Größe 1024, Deskriptorgröße 24×24 Pixel, Fensterschrittweite 2 Pixel in x- und y-Richtung, Fensterbreite 1 Deskriptorbreite bzw. 24 Pixel, BC-Distanz für das DTW und kein Anpassungsfenster.

Normalisierung	GW	JB-V
Ja	57,0	71,7
Nein	44,1	64,0

Tabelle A.0.3: Mean Average Precisions der Experimente zum Einfluss der DTW Normalisierung.

LGH Anpassung mit Otsu Binarisierung

Erste Experimente für eine Anpassung des LGH-Deskriptor auf den Schriftbereich zeigten schlechte Resultate. Grund dafür war eine hier ungeeignete Binarisierung mit dem Otsu-Schwellwert [Ots79].
Folgende Parameter waren für den JB-V Datensatz eingestellt: Fensterbreite 24 Pixel, 4×4 Zellen, L1-Distanz für das DTW, Anpassungsfenster 10%, $\mathtt{alpha} = -0,2$..

Distanz	LGH-Full	LGH-Otsu
L1	59,7	61,0

Tabelle A.0.4: Mean Average Precisions auf dem JB-V Datensatz.

VIN-Niblack verschiedene Fensterbreiten

Experimente zur Fensterbreite für die VIN-Niblack Repräsentation.
Folgende Parameter waren für den JB-V Datensatz eingestellt: L1-Distanz für das DTW, Anpassungsfenster 10%, $\mathtt{alpha} = -0,2$.
Folgende Parameter waren für den GW Datensatz eingestellt: L1-Distanz für das DTW, Anpassungsfenster 10%, $\mathtt{alpha} = -0,2$.

	GW	JB-V
Fensterbreite	4×4	5×8
12	-	-
16	49,3	67,8
20	28,0	-
24	31,2	**69,6**
28	32,4	-
32	-	67,2
40	**33,7**	-

Tabelle A.0.5: Mean Average Precisions für verschiedene Fensterbreiten der VIN-Niblack Sequenz.

LGH-Niblack verschiedene Anzahlen der Hauptorientierungen

Hier sind die Auswertungsergebnisse einiger Experimente zur Anzahl der Hauptorientierungen aufgeführt.
Folgende Parameter waren für den GW Datensatz eingestellt: $\alpha = -0.2$, Fensterschrittweite 5 Pixel, Fensterbreite 24 Pixel, 4×4 Zellen, L1-Distanz für das DTW, Anpassungsfenster 10%.
Folgende Parameter waren für den JB-V Datensatz eingestellt: $\alpha = -0.2$, Fensterschrittweite 2 Pixel, Fensterbreite 16 Pixel, 4×4 Zellen, L1-Distanz für das DTW, Anpassungsfenster 10%.

Anzahl	GW	JB-V
8	50,5	75,3
10	51,6	**75,0**
12	52,2	74,3
14	53,1	-
16	**53,7**	73,8

Tabelle A.0.6: Mean Average Precisions für verschiedene Anzahlen der Hauptorientierungen.

VIN-Niblack vs. LGH-Niblack mit 1 Hauptorientierung

Hier sind einige Auswertungsergebnisse zur VIN-Niblack Repräsentation und zur LGH-Niblack Repräsentation mit 1 Hauptorientierung (LGH-1-Niblack) aufgeführt.
Folgende Parameter waren für den GW Datensatz eingestellt: $\alpha = -0.2$, Fensterschrittweite 5 Pixel, Fensterbreite 40 Pixel, L1-Distanz für das DTW, Anpassungsfenster 10%.
Folgende Parameter waren für den JB-V Datensatz eingestellt: $\alpha = -0.2$, Fensterschrittweite 2 Pixel, Fensterbreite 24 Pixel, L1-Distanz für das DTW, Anpassungsfenster 10%.

	GW		JB-V	
Zelleneinteilung	VIN-Niblack	LGH-1-Niblack	VIN-Niblack	LGH-1-Niblack
4×4	33,7	33,7	66,3	59,9
5×8	35,1	34,3	69,6	63,6
4×20	35,9	34,7	-	-

Tabelle A.0.7: Mean Average Precisions für verschiedene Zelleneinteilungen.

VIN-Niblack verschiedene Zelleneinteilung

Auswertungsergebnisse zur VIN-Niblack Repräsentation in verschiedenen Zelleneinteilungen
Folgende Parameter waren für den GW Datensatz eingestellt: $\alpha = -0.2$, Fensterschrittweite 5 Pixel, Fensterbreite 40 Pixel, L1-Distanz für das DTW, Anpassungsfenster 10%.
Folgende Parameter waren für den JB-V Datensatz eingestellt: $\alpha = -0.2$, Fensterschrittweite 2 Pixel, Fensterbreite 24 Pixel, L1-Distanz für das DTW,

GW (Fensterbreite 40 Pixel)				JB-V (Fensterbreite 24 Pixel)			
	Spalten				Spalten		
Zeilen	8	10	20	Zeilen	6	8	12
3	34,5	34,7	35,0	4	68,1	68,5	68,4
4	35,3	35,6	**35,9**	5	69,3	**69,6**	69,0
5	35,1	35,2	35,6	6	68,4	68,7	68,8

Tabelle A.0.8: Mean Average Precisions für verschiedene Zelleneinteilungen der VIN-Niblack Repräsentation.

LITERATURVERZEICHNIS

[AFV13] Almazán, J. ; Fornés, A. ; Valveny, E.: Deformable HOG-Based Shape Descriptor. In: *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, 2013, S. 1022–1026

[AGFV14] Almazán, J. ; Gordo, A. ; Fornés, A. ; Valveny, E.: Segmentation-free word spotting with exemplar SVMs. In: *Pattern Recognition* 47 (2014), Nr. 12, S. 3967–3978

[AJ00] A.Vinciarelli ; J.Luettin: Off-Line Cursive Script Recognition based on continuous density HMM. In: *Proc. of International Workshop on Frontiers in Handwriting Recognition*, 2000, S. 493–498

[ARTL15] Aldavert, D. ; Rusiñol, M. ; Toledo, R. ; Lladós, J.: A study of Bag-of-Visual-Words representations for handwritten keyword spotting. In: *International Journal on Document Analysis and Recognition (IJDAR)* 18 (2015), Nr. 3, S. 223–234

[Con06] Congress, Library of: *Technical Standards for Digital Conversion of Text and Graphic Materials*. http://memory.loc.gov/ammem/about/techStandards.pdf. Version: Dezember 2006, Abruf: 31.03.2016

[DH73] Duda, Richard O. ; Hart, Peter E.: *Pattern Classification and Scene Analysis*. New York : Wiley, 1973. – 270–272 S.

[DT05] Dalal, N. ; Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* Bd. 1, 2005, S. 886–893

[FFMB12] Frinken, V. ; Fischer, A. ; Manmatha, R. ; Bunke, H.: A Novel Word Spotting Method Based on Recurrent Neural Networks. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 34 (2012), Nr. 2, S. 211–224

[Fin14] Fink, Gernot A.: *Markov Models for Pattern Recognition - From Theory to Applications*. 2. London : Springer, 2014

[FRG14] Fink, G. A. ; Rothacker, L. ; Grzeszick, R.: Grouping Historical Postcards Using Query-by-Example Word Spotting. In: *Proc. Int. Conf. on Frontiers in Handwriting Recognition*, 2014, S. 470–475

[GP09] Gatos, B. ; Pratikakis, I.: Segmentation-free Word Spotting in Historical Printed Documents. In: *Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference on*, 2009, S. 271–275

[Jä12] Jähne, Bernd: *Digitale Bildverarbeitung: und Bildgewinnung*. 7. Berlin, Heidelberg : Springer Berlin Heidelberg, 2012

[Low04] Lowe, D. G.: Distinctive Image Features from Scale-Invariant Keypoints. In: *International Journal of Computer Vision* 60 (2004), Nr. 2, S. 91–110

[LSP06] Lazebnik, S. ; Schmid, C. ; Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* Bd. 2, 2006, S. 2169–2178

[Man99] Manmatha, Nitin R.and S. R.and Srimal: Scale Space Technique for Word Segmentation in Handwritten Documents. In: *Proc. Scale-Space Theories in Computer Vision: Second International Conference*, 1999, S. 22–33

[MHR96] Manmatha, R. ; Han, Chengfeng ; Riseman, E.M.: Word spotting: a new approach to indexing handwriting. In: *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR '96, 1996 IEEE Computer Society Conference on*, 1996, S. 631–637

[Nib86] Niblack, Wayne: *An introduction to digital image processing*. Englewood Cliffs, N.J. : Prentice Hall, 1986. – 115–116 S.

[Nie03] Niemann, Heinrich: *Klassifikation von Mustern*. 2. 2003 http://www5.informatik.uni-erlangen.de/fileadmin/Persons/NiemannHeinrich/klassifikation-von-mustern/m00-www.pdf

[OD11] O'Hara, S. ; Draper, B. A.: Introduction to the Bag of Features Paradigm for Image Classification and Retrieval. In: *CoRR* abs/1101.3354 (2011)

[Ots79] Otsu, N.: A Threshold Selection Method from Gray-Level Histograms. In: *IEEE Transactions on Systems, Man, and Cybernetics* 9 (1979), Nr. 1, S. 62–66

[Pri15] Priese, Lutz: *Computer Vision - Einführung in die Verarbeitung und Analyse digitaler Bilder*. Berlin, Heidelberg : Springer Berlin Heidelberg, 2015

[PTV15] Puigcerver, J. ; Toselli, A. H. ; Vidal, E.: ICDAR2015 Competition on Keyword Spotting for Handwritten Documents. In: *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, 2015, S. 1176–1180

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

[RFM+15] Rothacker, L. ; Fisseler, D. ; Müller, G. G. ; Weichert, F. ; Fink, G. A.: Retrieving Cuneiform Structures in a Segmentation-free Word Spotting Framework. In: *Proc. of the Int. Workshop on Historical Document Imaging and Processing*, 2015, S. 129–136

[RM03] Rath, T. M. ; Manmatha, R.: *Word Image Matching Using Dynamic Time Warping*. 2003

[RM07] Rath, T. M. ; Manmatha, R.: Word spotting for historical documents. In: *International Journal of Document Analysis and Recognition (IJDAR)* 9 (2007), Nr. 2, S. 139–152

[RP08] Rodriguez, J. A. ; Perronnin, F.: Local gradient histogram features for word spotting in unconstrained handwritten documents. In: *Proc. of the 1st International Conference on Handwriting Recognition (ICFHR08)*, 2008, S. 7–12

[RP09] Rodríguez-Serrano, José A. ; Perronnin, Florent: Handwritten word-spotting using hidden Markov models and universal vocabularies. In: *Pattern Recognition* 42 (2009), Nr. 9, S. 2106–2116

[RRF13] Rothacker, L. ; Rusinol, M. ; Fink, G. 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

[RRLF14] Rothacker, L. ; Rusinol, M. ; Llados, J. ; Fink, G. A.: A Two-Stage Approach to Segmentation-Free Query-by-Example Word Spotting. In: *manuscript cultures* 1 (2014), Nr. 7, S. 47–57

[RVF12] Rothacker, L. ; Vajda, S. ; Fink, G. A.: Bag-of-Features Representations for Offline Handwriting Recognition Applied to Arabic Script. In: *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on*, 2012, S. 149–154

[SC78] Sakoe, H. ; Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. In: *Acoustics, Speech and Signal Processing, IEEE Transactions on* 26 (1978), Nr. 1, S. 43–49

[SC07] Salvador, S. ; Chan, P: Toward Accurate Dynamic Time Warping in Linear Time and Space. In: *Intelligent Data Analysis* 11 (2007), S. 561–580

[SF15] *Kapitel* A Modified Isomap Approach to Manifold Learning in Word Spotting. In: Sudholt, S. ; Fink, G. A.: *Pattern Recognition: 37th German Conference, GCPR 2015, Aachen, Germany, October 7-10, 2015, Proceedings*. Cham : Springer International Publishing, 2015, S. 529–539

[Sin01] Singhal, A.: Modern Information Retrieval: A Brief Overview. In: *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 24 (2001), Nr. 4, S. 35–42

[SP00] Sauvola, J. ; Pietikäinen, M.: Adaptive document image binarization. In: *Pattern Recognition* 33 (2000), S. 225–236

[SRF15] Sudholt, S. ; Rothacker, L. ; Fink, G. A.: Learning Local Image Descriptors for Word Spotting. In: *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, 2015, S. 651–655

[Tru07] True, N.: *Offline Word Spotting in Handwritten Documents*. https://cseweb.ucsd.edu/classes/fa07/cse252c/projects/ntrue.pdf. Version: 2007, Abruf: 08.04.2016

[TT09] Terasawa, K. ; Tanaka, Y.: Slit Style HOG Feature for Document Image Word Spotting. In: *Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference on*, 2009, S. 116–120

[ZPG14] Zagoris, K. ; Pratikakis, I. ; Gatos, B.: Segmentation-Based Historical Handwritten Word Spotting Using Document-Specific Local Features. In: *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, 2014, S. 9–14