technische universität
dortmund

# Textdetektion und Indizierung in historischen Dokumenten

**Bachelorarbeit**
**(Überarbeitete Fassung)**

**Keng Hong Chai**
**27. November 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

Die automatische Analyse von historischen handgeschriebenen Dokumenten gilt als noch nicht vollständig gelöst. Somit ist die Dokumentenanalyse ein aktuelles Thema in der Mustererkennung [RVF12, RRF13, RATL15]. Im Gegensatz zu modernen Dokumenten, in denen Ansätze auf Basis von Optical Character Recognition (OCR) schon nahezu optimale Ergebnisse liefern, sind die bisherigen Ergebnisse auf historischen Dokumenten noch nicht ausreichend. Das Ziel der OCR ist es in Bildern auftretende Wörter zu erkennen und in maschinenkodierte Wörter umzuwandeln. Dieser Prozess der Zuweisung eines maschinenkodierten Wortes zu einem Wortabbild wird auch Transkription genannt.

Im Bereich der modernen Dokumente, in denen ein einheitlicher Schriftsatz vorhanden ist und sowohl der Text als auch der Hintergrund klar voneinander trennbar sind, funktioniert das Verfahren sehr gut (vgl. [Bor14]). Abbildung 1.0.1 verdeutlicht die verschiedenen Rahmenbedingungen der beiden Problemdomänen und zeigt, dass diese Annahmen nicht ohne weiteres auf historische Dokumente übertragbar sind.

Anders als bei maschinengeschriebenen Dokumenten ist der Text in historischen Dokumenten handgeschrieben. Die Analyse von handgeschriebenen Texten beinhaltet mehrere Schwierigkeiten. Ein großes Problem ist die mögliche Variabilität in der Schreibweise von Wörtern. Selbst Vorkommen des gleichen Wortes können sich stark voneinander unterscheiden. In Abbildung 1.0.2 sind verschiedene Vorkommen des Wortes „and“ dargestellt, in denen starke Unterschiede in der visuellen Erscheinung zwischen dem Buchstaben „d“ existieren.

Eine weitere Schwierigkeit ist der tendentiell schlechte Qualitätszustand der Dokumente, Effekte wie Vergilbung entstehen, sodass selbst menschlichen Lesern das Erkennen der Wörter schwer fällt. Weitere Probleme die entstehen können, sind das Verwischen von Tinte und Bleed-Through, das Durchsickern von Tinte auf die Rückseite des Dokumentes und führen zu ungewünschter Variabilität im Schriftbild. In Abbildung 1.0.3 ist eine Dokumentenseite abgedruckt, in der die Qualität beeinträchtigt ist und einige Wörter deshalb nur noch schwer erkennbar sind.

Diese Bedingungen erschweren die automatische Transkription von historischen handgeschriebenen Dokumenten. Dennoch gilt es diesen Schritt zu automatisieren, da für größere Dokumentensammlungen, wie zum Beispiel der George Washington

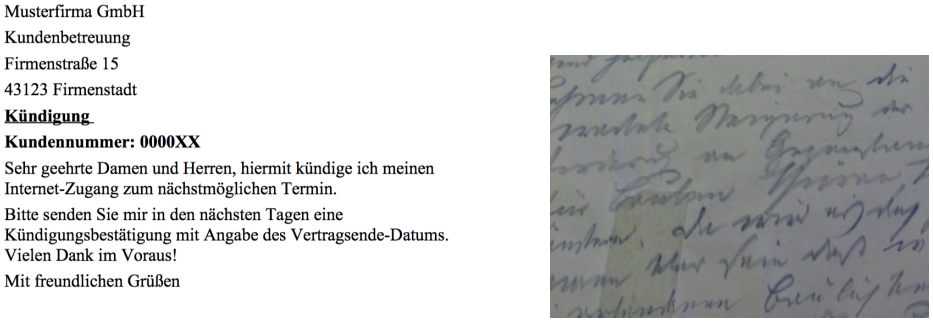



Abbildung 1.0.1: Links befindet sich ein Ausschnitt eines modernen Dokumentes. Rechts ist ein Ausschnitt eines historisches Dokumentes abgebildet.

Abbildung 1.0.2: Zwei unterschiedliche Vorkommen des Wortes „and“. Auffällig ist der Unterschied der Schreibweise des „d“.

Dokumentensammlung[1], eine manuelle Transkription ineffizient ist. Intuitiv können hierfür holistische Worterkenner verwendet werden, die als Ziel haben die Wortabbilder eines Dokumentes zu erkennen. Bislang entstehen über diese Verfahren jedoch noch viele Fehlerfälle [RRF13, RM06].

Eine andere Möglichkeit dafür bietet die Indizierung, welche sich stattdessen mit der visuellen Ähnlichkeitsbewertung von Wortabbildern befasst. Die Aufgabe dieser Bachelorarbeit soll es sein wichtige historische Dokumente wie zum Beispiel Briefe, Verträge und Abkommen zu indizieren, mit dem Ziel größere Dokumentensammlungen zu transkribieren. Über den Indexeintrag soll es zusätzlich Nutzern ermöglicht werden, alle weiteren Vorkommen des Wortes aufzufinden, analog zu einer modernen Textsuche in Editoren. Diese Aufgabe lässt sich in zwei Bereiche einteilen: Das Detektieren der Wortabbilder des Dokumentes und der darauf aufbauenden Indizierung. Nach bestem Wissen und Gewissen des Autors ist dies die erste Arbeit, die beide Gebiete behandelt. Bisherige Arbeiten hierzu funktionieren segmentierungsbasiert und

1 Die George Washington Dokumentensammlung im Library of Congress besteht aus schätzungsweise 65000 Dokumenten (https://memory.loc.gov/ammem/gwhtml/).

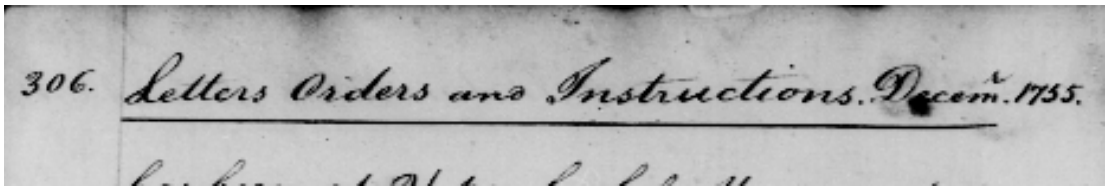



Abbildung 1.0.3: In der Abbildung ist ein Dokument darstellt, dessen Qualität durch Alterungseffekte stark beeinträchtigt ist.

setzen in einem vorherigen Schritt lokalisierte Wortabbilder voraus. Im Folgenden werden die zwei Hauptbereiche der Dokumentenindizierung vorgestellt.

## 1.1 TEXTDETEKTION

Ein wichtiger Schritt der Dokumentenverarbeitung ist die Textdetektion. Der Bereich der Textdetektion beschäftigt sich mit dem Auffinden von Wörtern in Bildern. Zu Beginn werden die Dokumente digitalisiert, indem diese gescannt werden. In diesen Bildern gilt es die einzelnen Wörter der Dokumente zu finden. Dafür werden die Bilder in Ausschnitte unterteilt, welche die einzelnen Wörter enthalten. Dieser Schritt wird auch Segmentierung genannt. Das Ziel der Bildsegmentierung ist es, inhaltlich relevante Objekte in Bildern zu extrahieren, die im Szenario der historischen Dokumente den einzelnen Wortabbildern entsprechen. Gleichzeitig sollen Teile des Bildes, welche für die Historiker relevanten Szenarios nicht von Bedeutung sind, wie zum Beispiel Hintergründe und Artefakte verworfen werden. In der Praxis werden die gefundenen Wortabbilder dann mit Rechtecken, den sogenannten Bounding Boxes, versehen mit dem Ziel die Wörter dabei bestmöglich einzugrenzen. Für kleinere Teilmengen von Dokumentensammlungen existieren manuell angefertige Annotationen mit Koordinaten der Wortabbilder und jeweiliger Transkription. In Abbildung 1.1.1 ist ein Ausschnitt eines Dokumentenabbildes dargestellt, in dem das Dokument mithilfe der Annotation segmentiert wurde. Gegenwärtig lassen sich die aktuellen Verfahren in zwei Kategorien unterteilen:

- **Gleitendes Fenster:** Diese Verfahren arbeiten mit einem Fenster, welches über das Bild bewegt wird [WBB11, CCC+11]. In diesem Fenster wird geprüft, ob der Bildbereich innerhalb des Fensters mit einem Wort korrespondiert. Damit solch ein System möglichst viele Wörter so genau wie möglich detektiert, ist es nötig das Verfahren mit vielen unterschiedlichen Fenstergrößen zu wiederholen. Dadurch kann die Komplexität stark ansteigen (vgl. [YD15]).

Abbildung 1.1.1: Die Bounding Boxes wurden hier manuell erstellt, um eine optimale Segmentierung zu zeigen.

- **Kombinieren von Textregionen** Ein weiteres oft genutztes Verfahren basiert auf dem Lokalisieren von möglichen Textbereichen. Diese Bereiche werden dann miteinander kombiniert um Worthypothesen zu erstellen. Für jede der eventuellen Textkandidaten gilt es nun, diese in Text bzw. Hintergrund zu klassifizieren. Eine erfolgreiche Methode zum Auffinden von Textbereichen ist der Maximally Stable Extremal Regions Detektor [NM11a, QM16, TD15].

## 1.2 INDIZIERUNG

Die Aufgabe der Textindizierung ist es, analog zu Stichwortverzeichnissen in modernen Büchern, einen Index für eine Dokumentensammlung zu erstellen, der die wichtigen Wörter des Textes enthalten soll. Menschlichen Nutzern soll dann ermöglicht werden, über den jeweiligen Registereintrag alle Vorkommen des Stichwortes aufrufen zu können. Die grundliegende Idee dahinter ist es Gruppen zu bilden, die jeweils alle Vorkommen eines bestimmten Wortes zusammen zu fassen. Unter der Annahme, dass zwischen Vorkommen des gleichen Wortes visuelle Gemeinsamkeiten bestehen, ist es möglich mithilfe von Clustering ähnliche Wörter zusammen zu fassen. Diese Idee basiert auf der Arbeit von Manmatha et al. [MHR96]. Im optimalen Fall entsteht für jedes verschiedene Wort in der Dokumentensammlung ein Cluster, in dem alle Vorkommen des jeweiligen Wortes enthalten sind. Für jedes Cluster wird eine repräsentative Annotation vergeben, welches an alle weiteren Wortabbilder innerhalb des Clusters propagiert wird. Hierdurch wird eine komplette Transkription der Dokumentensammlung generiert.

Wird nur ein Stichwortverzeichnis statt einer kompletten Transkription verlangt, können Cluster von nicht aussagekräftigen Wörtern anhand von Heuristiken verworfen werden. Zu oft auftauchende Wörter entsprechen mit hoher Wahrscheinlichkeit Stoppwörtern, die keinen semantischen Wert für den Text haben. Tendentiell sind also Cluster relevant, die aus nicht zu vielen Wortabbildern bestehen [MHR96, RM06, Luh58, Pia14].

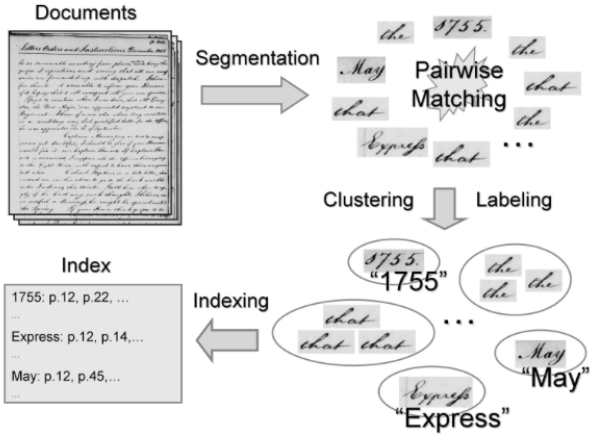



Abbildung 1.2.1: Der generelle Ablauf des Verfahrens. Nach dem Scannen der Dokumente werden die Wortabbilder segmentiert. Für Wortabbilder wird eine geeignete Merkmalsrepräsentation berechnet und mithilfe von Clustering gruppiert. Für jedes Cluster wird ein repräsentatives Label bestimmt, welches an alle weiteren Clustervektoren propagiert wird. Hierdurch wird die Dokumentensammlung transkribiert. Aus den entstandenen Clustern können dann die inhaltlich relevanten Registereinträge gewählt. Entnommen aus [RM06].

Das Ziel der vorliegenden Arbeit ist es, ein Verfahren für die bisher unbehandelte segmentierungsfreie Indizierung zu erstellen. In Abbildung 1.2.1 ist der vollständige Ablauf eines solchen Systems dargestellt. Die Segmentierung extrahiert die Wörter der Dokumente. Hierbei wird der MSER Detektor [MCUP02] verwendet, der nach bisherigem Kenntnisstand des Autors noch nicht im Bereich der historischen Dokumente zum benutzt wurde. Für die gefundenen Wortabbilder gilt es eine geeignete charakteristische Merkmalsrepräsentation zu finden. Diese werden dann mithilfe von Clustering gruppiert und aus den generierten Clustern die Registereinträge gewählt. Hierfür wird der Spherical K-Means [DM01, BDGS05] benutzt, der nach aktuellem Wissensstand des Autors für diese Aufgabe noch keine Anwendung fand.

## 1.3 GLIEDERUNG DER ARBEIT

Der Aufbau der Bachelorarbeit gestaltet sich wie folgt. In Kapitel 2 werden grundlegende Methodiken erläutert, die dem weiteren Verständnis dieser Arbeit dienen. Kapitel 3 beschäftigt sich mit den bisher veröffentlichen Arbeiten der Dokumentenanalyse, die einen großen Bezug zur vorliegenden Bachelorarbeit haben. Kapitel 4 be-

schreibt das komplette Verfahren, welches für die genannten Aufgaben erstellt wurde. Kapitel 5 erläutert die verschiedenen Experimente und die Datensätze auf denen evaluiert wurde. In Kapitel 6 werden am Ende die Ergebnisse zusammengefasst, um ein abschließendes Fazit zu geben.

# 2 GRUNDLAGENKAPITEL

Das folgende Kapitel behandelt einige grundlegende Verfahren, die für das weitere Verständnis der Arbeit nötig sind. Dabei bieten die Abschnitte einen Einblick in die wesentlichen Aspekte der Konzepte. Detaillierte Erklärungen sind in der jeweils angegeben Literatur zu finden.

In Kapitel 1 wurde die Indizierung von historischen Dokumenten im segmentierungsfreien Kontext beschrieben. Bisherige Verfahren zur Indizierung lassen sich in den segmentierungsbasierten Fall einordnen. Somit ist die vorliegende Arbeit nach bestem Wissen und Gewissen des Autors die erste, welche die segmentierungsfreie Indizierung untersucht. Da in diesem Fall von keiner vorhandenen, optimalen Segmentierung ausgegangen wird, ist ein Schritt zur Textdetektion nötig. Wesentlicher Bestandteil der Textdetektion wird der MSER Detektor sein, der in Abschnitt 2.1 erklärt wird und zum Auffinden von möglichen Textbereichen verwendet wird. Der MSER Detektor hat sich für Bilder von realen Szenen bewährt [NM11a, QM16, TD15]. Im Bereich der historischen Dokumente wurde der MSER Detektor nach bestem Wissen und Gewissen des Autors bislang noch nicht verwendet.

Für Wortabbilder gilt es charakteristische Merkmalsrepräsentationen zu finden, damit klare Differenzierungen zwischen unterschiedlichen Wörtern möglich sind. Hierfür bedient sich die Arbeit einer Merkmalsdarstellung von Wortabbildern, die sich im Bereich der Dokumentenanalyse etabliert hat. Bestehend aus dem SIFT Deskriptor (Abschnitt 2.2) und der darauf aufbauenden Bag of Features Repräsentation (Abschnitt 2.3) werden Wortabbilder als Spatial Pyramid repräsentiert.

Abschnitt 2.4 beschreibt das Clustering, welches zwei Anwendungsgebiete in der vorliegenden Arbeit hat. Einerseits wird es für die Merkmalsberechnung von Bag of Features Darstellungen benötigt. Im vorherigen Kapitel wurde zudem die grundlegende Verfahrensweise der Indizierung eingeführt. Mit dem Ziel ähnliche Wörter zu gruppieren wird hier ein separates Clustering durchgeführt. Insbesondere wird hierfür der Spherical K-Means eingeführt, der nach aktuellem Kenntnisstand des Autors zum bisherigen Zeitpunkt für diesen Zweck noch nicht verwendet wurde.

## 2.1 MAXIMALLY STABLE EXTREMAL REGIONS DETEKTOR

Der Maximally Stable Extremal Regions Detektor (MSER) wurde von Matas et al. in [MCUP02] vorgestellt und wird benutzt, um homogene Bildbereiche zu finden. Diese Bildregionen sind die sogenannten MSER, die aus Pixeln bestehen, die sich bezüglich ihrer Grauwertintensität homogen verhalten. Der MSER Detektor wird aufgrund dieser Eigenschaft zur Lokalisierung von Textregionen genutzt. Da Schriftzüge in der Regel aus Bildpixeln ähnlicher Intensitäten bestehen, korrespondieren generierte MSER mit möglichen Textbereichen [NM11a, TD15, CTS+11, NM11b, YD15]. Abbildung 2.1.1 bekräftigt diese Aussage. In diesem Bild wurde der MSER Detektor angewendet und die berechneten Regionen mit Rechtecken umrahmt. Die Textstellen werden durch die MSER abgedeckt. Für das MSER Verfahren muss vorab der Begriff des Binärbildes erläutert werden.

### 2.1.1 *Binärbild*

Die Binarisierung eines Graustufenbildes ist ein Verfahren, welches ein Binärbild generiert, um die Bildpixel in Hintergrund- und Objektpixel zu unterteilen (vgl. [NFHS11] Kapitel 4.5). Im Kontext von Dokumenten entsprechen Objekte den Wörtern des Dokumentes. Das Binärbild wird mithilfe eines Schwellwertverfahren berechnet. Dafür werden alle Pixel mit einem Schwellwert $t$ verglichen und die Pixelwerte mit Gleichung 2.1.1 neu berechnet. Pixelwerte, die größer bzw. gleich $t$ sind, werden auf 1 und Pixel kleiner als $t$ auf 0 gesetzt (vgl. [NFHS11] Kapitel 4.5). Visuell entsprechen Pixel mit dem Wert 0 schwarzen Bildpunkten (Schrift) und Pixel mit dem Wert 1 weißen Bildpunkten (Hintergrund).

$$B(x) = \begin{cases} 1, & \text{falls } x \geqslant t \\ 0, & \text{sonst} \end{cases} \tag{2.1.1}$$

Mithilfe der Einteilung in Hintergrund und Schrift ist prinzipiell eine Segmentierung möglich. Ein großes Problem hierbei ist die Wahl des Schwellwertes $t$. Dieser müsste so festgelegt werden, dass eine klare Trennung von Schrift- und Hintergrundpixeln möglich ist. In historischen Dokumenten können Effekte, wie Verwischen und Verbleichen von Schriftverläufen und Vergilbung, auftreten. Durch die feste Wahl auf genau einen Wert für $t$ können diese Effekte zu schlechteren Ergebnissen führen.

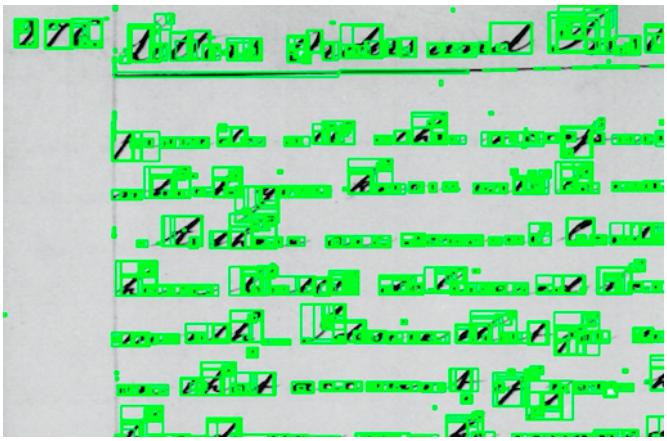


Abbildung 2.1.1: Generierte MSER Regionen auf einem Ausschnitt des Washington Datensatzes. Zu beachten ist hierbei, dass die umrahmenden Rechtecke der einzelnen MSER zu sehen sind.

### 2.1.2 *Zusammenhangskomponenten*

Auf einem Binärbild lassen sich Zusammenhangskomponenten (Blobs) definieren. Zuvor muss dafür der Begriff Pixel Nachbarschaft erläutert werden. Unterschieden wird üblicherweise zwischen der 4 und 8 Nachbarschaft. In der 4 Nachbarschaft eines Pixels $p$ liegen die Pixel überhalb, unterhalb, rechts und links von $p$. Zusätzlich werden in der 8 Nachbarschaft die diagonal angrenzenden Pixel betrachtet (vgl. [GW06] Seite 90 ff.).

Mithilfe der Nachbarschaftsdefinition sind jetzt Zusammenhangskomponenten berechenbar. Dafür wird eine Menge $V$ von Pixelwerten definiert, für die Zusammenhangskomponenten berechnet werden sollen. Die restlichen Pixelwerte werden als Hintergrund festgelegt. Im Fall von Binärbildern, in denen Pixelwerte entweder den Wert 0 oder 1 annehmen können, wird die Menge $V$ entweder auf $V = \{0\}$ (schwarze Zusammenhangskomponenten) oder $V = \{1\}$ (weiße) gesetzt. Ausgehend von einem Startpixel $p$ mit einem Wert aus $V$ werden die Pixel der verwendeten Nachbarschaftsdefinition betrachtet. Die Nachbarn, die ebenfalls einen Wert aus $V$ besitzen, werden der Zusammenhangskomponente hinzugefügt. Dies wird für die neu hinzugekommenen Pixel wiederholt, bis diese nicht weiter wächst (vgl. [GW06] Seite 90 ff.).

Ein Algorithmus zur Bestimmung aller Zusammenhangskomponenten eines Bildes berechnet iterativ alle maximalen Zusammenhangskomponenten (vgl. [GW06] Seite

667-669). Prinzipiell lässt sich dieses Verfahren auch auf Graustufenbilder übertragen, indem Intervalle von Graustufen für die Menge V betrachtet werden.

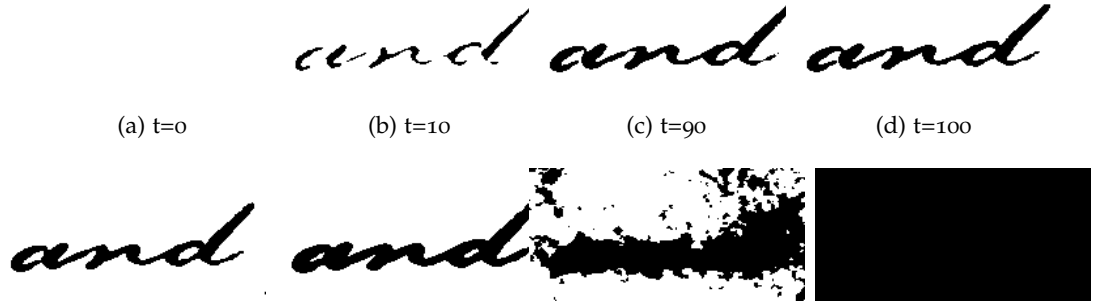
### 2.1.3 *Berechnen von MSER Regionen*

Mit den oben erklärten Konzepten kann das MSER Verfahren beschrieben werden. Die Menge der MSER eines Graustufenbildes berechnet sich wie folgt: Anhand aller möglichen Schwellwerte von 0...255 wird das Bild binarisiert. Auf den dadurch generierten Binärbildern werden die Zusammenhangskomponenten berechnet. Dabei wurde in der Arbeit von Matas et al., welche den MSER Detektor vorgestellt hat, die 4 Nachbarschaft verwendet [MCUP02]. Aufgrund der Eigenschaft, dass alle Randpixel außerhalb der Zusammenhangskomponente ungleich der Pixelwerte innerhalb des Blobs sind, werden diese auch Extremal Regions genannt. Abbildung 2.1.2 zeigt diesen Verlauf exemplarisch mit einigen ausgewählten Schwellwerten. Das erste Binärbild ist komplett weiß. Bei größer werdenden Schwellwerten entstehen schwarze Zusammenhangskomponenten, die zunehmend an Größe gewinnen. Dabei verschmelzen die Blobs, bis am Ende ein komplett schwarzes Bild entsteht.

Für Blobs lässt sich jetzt ein Kriterium definieren, welches die Stabilität einer Komponente bezüglich mehrerer Schwellwerte beschreibt. Die Anzahl von Schwellwerten, die zusätzlich betrachtet werden, wird als Parameter $\Delta$ bezeichnet. Ändert sich ein Blob über mehrere Schwellwerte kaum bis gar nicht, gilt dieses als stabil. Dabei wird die Fläche eines Blobs betrachtet, die der Anzahl seiner Pixel entspricht. Eine Region zu einem Schwellwert $t_i$ ist dann maximal stabil, wenn der Flächenunterschied über die von $\Delta$ definierte Folge von Schwellwerten ein lokales Minimum bildet [MCUP02]. Diese Teilmenge der Extremal Regions werden aufgrund der genannten Eigenschaften Maximally Stable Extremal Regions genannt und sind die detektierten Bildbereiche [MCUP02].

## 2.2 SIFT VERFAHREN

Ein lokaler Bilddeskriptor, der sich im Bereich der Dokumentenanalyse etabliert hat, ist der Scale Invariant Feature Transform (SIFT) Deskriptor [RRF13, RVF12, RATL15, RATL11, RF15]. Lokale Bilddeskriptoren zeichnen sich dadurch aus, dass zu einem Bildpunkt zusätzlich die lokale Umgebung in die Berechnung der Deskriptoren miteinfliesst. Zusammen mit dem SIFT Detektor stellt Lowe ein Verfahren zur Beschreibung von auffälligen Bildbereichen zur Verfügung [Low04].



(a) t=0 (b) t=10 (c) t=90 (d) t=100

(e) t=110 (f) t=170 (g) t=200 (h) t=220

Abbildung 2.1.2: Das MSER Verfahren binarisiert das Bild für jeden möglichen Schwellwert. Zu Beginn ($t = 0$) ist das Bild weiß, denn alle Pixelwerte sind größer bzw. gleich 0. Mit größeren Schwellwerten t entstehen schwarze Blobs, die mit steigendem t wachsen und miteinander verschmelzen. Am Ende sind alle Pixelwerte kleiner als der aktuelle Schwellwert und das Bild ist komplett schwarz. Die MSER sind die Bereiche, die für eine Reihe von Schwellwerten konstant und stabil bleiben.

### 2.2.1 *Dense Grid*

Ursprünglich wird im SIFT Verfahren eine Keypoint Detektion verwendet, um besonders auffällige Bildbereiche zu identifizieren. Der Detektor wird in der Dokumentenanalyse üblicherweise vernachlässigt und ein Dense Grid Verfahren genutzt [RATL11, RATL15, RRF13]. Hierfür wird, wie in Abbildung 2.2.1 dargestellt, ein regelmäßiges Gitter über das Bild gelegt und an allen Gitterpunkten des Gitters ein Deskriptor berechnet.

### 2.2.2 *Gradientenbild*

Bevor der SIFT Deskriptor im Weiteren beschrieben wird, muss zuvor der Begriff des Gradientenbildes eingeführt werden. Der Gradient eines Bildes besteht aus den partiellen Ableitungen erster Ordnung des Bildes (vgl. [GW06] Seite 187 ff.). Dabei werden im zweidimensionalen Fall die Ableitungen in x- und y-Richtung betrachtet.

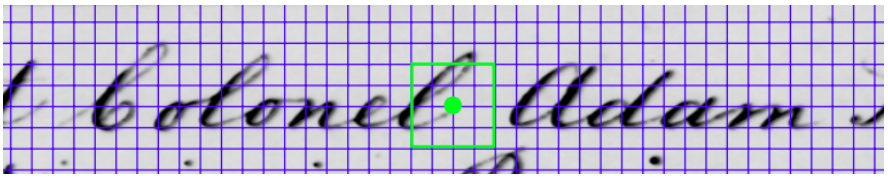


Abbildung 2.2.1: Berechnung der Deskriptoren in einem dichten Gitter. An jedem Gitterpunkt wird ein Deskriptor berechnet. Abbildung modelliert nach [RF16].

Die Ableitungen können mit dem Sobel Operator angenähert werden. Dieser besteht aus den folgenden Masken (vgl. [GW06] Seite 188-189.):

$$M_x = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}, \qquad M_y = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} \tag{2.2.1}$$

Ein Bild kann auch als diskretes zweidimensionales Signal interpretiert werden. Dann lassen sich die partiellen Ableitungen durch diskrete Faltungen des Bildes mit den Masken $M_x$ und $M_y$ berechnen (vgl. [NFHS11] Kapitel 5.4). Bei einer Faltung eines Bildes mit einer Maske wird ein neues Bild erzeugt, indem die Werte der Pixel neu berechnet werden. Jeder Bildpunkt wird durch eine Addition mit seinen gewichteten Nachbarpixeln gebildet. Dabei sind die zu betrachtenden Nachbarpixel und deren Gewichtung durch die Maske gegeben (vgl. [NFHS11] Kapitel 5.2). Die Faltungen mit den Masken $M_x$ und $M_y$ erzeugen den Gradienten in x-Richtung ($G_x$) bzw. y-Richtung ($G_y$). Der Gradient G ist dann definiert durch:

$$G = \begin{pmatrix} G_x \\ G_y \end{pmatrix} \tag{2.2.2}$$

Extremalstellen von $G_x$ und $G_y$ korrespondieren mit Kanten (vgl. [Jä12] Kapitel 12.4). Mit $G_x$ und $G_y$ können also horizontale bzw. vertikale Kanten gefunden werden. Die Berechnung eines Gradienten ist somit ein geeignetes Verfahren, um Schriftverläufe zu charakterisieren.

Für die Berechnung von SIFT Deskriptoren werden die Gradientenmagnituden $G_{mag}$ und Gradientenorientierungen $G_{dir}$ benötigt und berechnen sich wie folgt:

$$G_{mag} = \sqrt{G_x^2 + G_y^2}, \qquad G_{dir} = \arctan\left(\frac{G_y}{G_x}\right) \tag{2.2.3}$$

Die Magnituden eines Gradienten beschreiben also deren Länge, während die Orientierungen ihre Richtungen angeben (vgl. [NFHS11] Kapitel 5.4).

### 2.2.3 *SIFT Deskriptor*

An den Gitterpunkten werden jeweils Deskriptoren berechnet. Dabei dienen die Gitterpunkte als Mittelpunkte für die Deskriptoren. Für die Berechnung eines SIFT Deskriptor wird um den Mittelpunkt ein konfigurierbares n x n Fenster erstellt. In jeder dieser Zellen des Fensters wird auf Basis des Gradientenbildes ein Histogramm berechnet. Histogramme bestimmen die Häufigkeiten über Vorkommen von in Klassen (engl. Bins) eingeteilten Daten (vgl. [Ste03] Seite 66-67). Die Klassen des Histogramms entsprechen den auf die acht Hauptrichtungen quantisierten Gradientenorientierungen. Anstatt die absoluten Häufigkeiten der Vorkommen der Klassen zu zählen, bestimmen die Magnituden den Anteil zur entsprechenden Klasse des Histogramms. Die Gradientenmagnituden werden zusätzlich mit einem Gaußfenster gewichtet, sodass Magnituden, die vom Mittelpunkt weiter entfernt sind, weniger ins Gewicht fallen.

Das Fenster wird in der Regel in 4 x 4 Zellen unterteilt, womit dann insgesamt 4 x 4 = 16 Histogramme mit jeweils 8 Klassen entstehen. Durch das Konkatenieren der Histogramme ergibt sich der 16 x 8 = 128 dimensionale Deskriptor [Low04]. Abbildung 2.2.2 visualisiert die Berechnung des SIFT Deskriptors beispielhaft für eine 2 x 2 Zellenunterteilung. Im linken Teil der Abbildung werden die Orientierungen und Magnituden dargestellt. Der rechte Teil visualisiert die entsprechenden Gradientenhistogramme.

## 2.3 BAG OF FEATURES

Um die Ähnlichkeit zwischen Wortabbilder bewerten zu können, gilt es eine geeignete Merkmalsdarstellung zu finden mit der visuell ähnliche Wörter annähernd gleich beschrieben werden. Zusätzlich sollen unterschiedliche Wörter unähnliche Repräsentationen bekommen. Der folgende Abschnitt erklärt das Bag of Features (BoF), welches sich hierfür in der Dokumentenanalyse bewährt hat [RVF12, RRF13, RF15, RATL15].

Das Bag of Features Modell hat seinen Ursprung im Bag of Words Konzept. Das Bag of Words Modell wird in der Analyse maschinenlesbarer Texte benutzt. Unter der Annahme, dass die Wörter des Textes charakteristische Merkmale sind, wird eine Bag of Words Repräsentation berechnet, indem ein Histogramm über die Vorkommen der

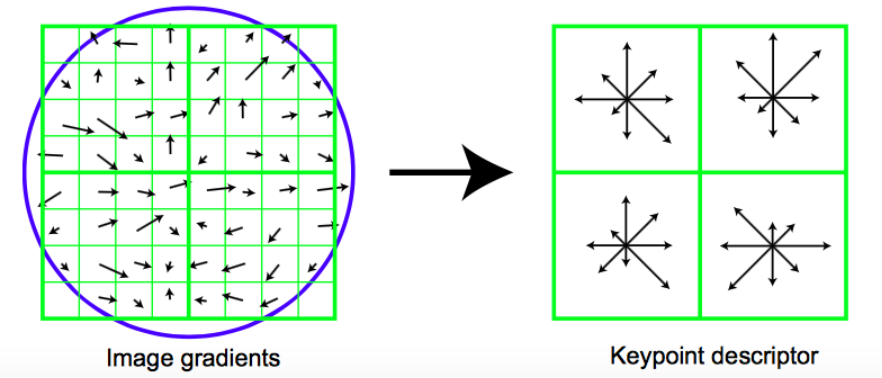



Abbildung 2.2.2: Links: Fensteraufteilung in 2 x 2 Zellen um den Deskriptor. Die Orientierungen sind als Pfeilrichtungen dargestellt und die Magnituden ensprechen den Pfeillängen. Das Gaußfenster mit dem die Magnituden gewichtet werden ist als blauer Kreis dargestellt.
Rechts: Für jede der einzelnen Zellen wird ein Histogramm der Orientierungen berechnet. Dabei werden die Orientierungen auf die acht Hauptrichtungen quantisiert und die jeweilige Magnitude bestimmt den Anteil zum Histogrammeintrag. Bild entnommen aus [Low04].

verschiedenen Wörter im Text gebildet wird. In der Regel werden in diesem Schritt nur die Wörter aus einem vorher definierten Vokabular betrachtet. Somit ergibt sich der sogenannte Termvektor, in dem jeder Eintrag die Vorkommen eines Wortes aus dem Vokabular zählt (vgl. [OD11]).

Dieses Konzept kann auf allgemeine Bilder übertragen werden, indem von der Wahl der Merkmale abstrahiert wird. Anstatt ein Histogramm über Wörter eines Textes zu bilden, wird ein Histogramm über Bilddeskriptoren berechnet. Ein hierfür oft genutzter Deskriptor ist der in Abschnitt 2.2 beschriebene SIFT Deskriptor [OD11, RATL15, RATL11, RF15, RRF13, RVF12]. Parallel zum Bag of Words wird ein Vokabular generiert, welches in Anlehnung dessen visuelles Vokabular genannt wird. Dafür werden auf einem Beispieldatensatz Deskriptoren berechnet, die dann geclustert werden (Näheres in Abschnitt 2.4). Das generierte Kodebuch bestehend aus den jeweiligen Clusterzentroiden entspricht dem visuellen Vokabular. Analog zum Bag of Words Ansatz werden die Kodes auch visuelle Wörter genannt. Eine Bag of Features Repräsentation berechnet sich dann, indem das Histogramm der anhand des Kodebuches quantisierten Deskriptoren des Bildes bestimmt wird (vgl. [OD11]). Abbildung 2.3.1 visualisiert den Ablauf einer Bag of Features Berechnung.

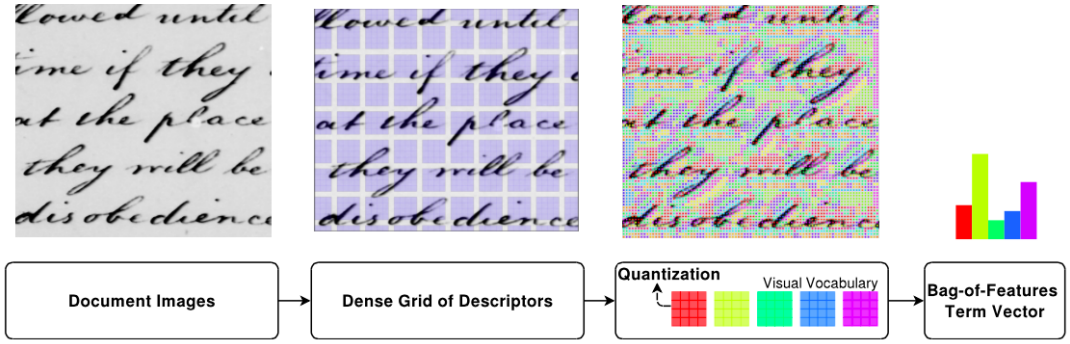



Abbildung 2.3.1: Für die Bag of Features Berechnung werden SIFT Deskriptoren in einem dichtem Gitter berechnet (blaue Quadrate). Diese werden anhand des visuelles Vokabular quantisiert (bunte Punkte). Schließlich wird der Termvektor gebildet, der die Vorkommen der quantisierten Deskriptoren zählt. Abbildung entnommen aus [FR14].

Durch solch eine Bag of Features Modellierung ist es zwar möglich Aussagen über die Frequenzen von bestimmten visuellen Wörtern zu tätigen, bei der ungeordneten Darstellung eines Bildes als Histogramm ist jedoch die räumliche Anordnung der visuellen Wörter nicht mehr rekonstruierbar.

Lazebnik et al. stellten aus diesem Grund den Beyond Bag of Features Ansatz vor, der das Bag of Features Modell mit einer Spatial Pyramid kombiniert, um diesem räumliche Informationen über die Vorkommen der Deskriptoren hinzuzufügen [LSP06]. Anfangs wurde das Verfahren benutzt, um Bildklassifizierungen durchzuführen. Im Bereich des Word Spotting wird die Spatial Pyramid Repräsentation für einzelne Wortbilder berechnet. Verschiedene Arbeiten haben gezeigt, dass die Bag of Features mit der Spatial Pyramid zu besseren Ergebnissen führt [LSP06, RATL15, RATL11].

In Abbildung 2.3.2 ist die Berechnung einer Spatial Pyramid Repräsentation eines Wortabbildes dargestellt. Das Wortabbild wird in immer feinere Teilbilder unterteilt, in denen jeweils lokale Bag of Features Histogramme gebildet werden. In dem Beispiel ist das Wortbild in zwei weitere Teilbilder unterteilt worden. Zusätzlich zur globalen BoF Darstellung wird für das linke, als auch rechte Teilbild ein lokales Histogramm bestimmt. Durch das Konkatenieren der berechneten Histogramme entsteht die Spatial Pyramid, womit sehr hochdimensionale Vektoren generiert werden. Mithilfe der Einteilung in Subregionen ist es möglich, die relative Position eines Deskrip-

tors innerhalb eines Bildes einzugrenzen und vermeidet somit den Verlust jeglicher Informationen über die räumliche Anordnung der Deskriptoren.

## 2.4 CLUSTERING

Die Clusteranalyse hat als Ziel ähnliche Datenpunkte (Vektoren) einer große Datenmenge in Partitionen (Cluster) einzuteilen. Es handelt sich hierbei um ein unüberwachtes Verfahren, dass ohne annotierte Daten arbeitet. Im Gegensatz hierzu stehen überwachte Lernverfahren in denen, mithilfe von Beispieldaten, für die eine korrekte Klassifikation bekannt ist, ein Modell trainiert wird. Unter der Annahme, dass Vektoren innerhalb eines Clusters ähnlich sind, wird für diese ein gemeinsamer Repräsentant ermittelt. Die Repräsentanten (Zentroiden) der Cluster ergeben das Kodebuch, mit dem eine spätere Vektorquantisierung möglich ist (vgl. [Fin14] Kapitel 4).

### 2.4.1 *Lloyd Algorithmus*

Zur Generierung eines Kodebuches für die in Abschnitt 2.3 beschriebene Bag of Features wird der Lloyd Algorithmus genutzt. Der Lloyd Algorithmus ist ein iteratives Verfahren, das in jedem Iterationsschritt eine lokale optimierte Partitionierung bestimmt. Sei $S = \{x_1, x_2, ..., x_n\}$ die Menge der Eingabevektoren. Aus der Eingabemenge S werden zufällig $k$ Vektoren gewählt. Das $k$ ist hierbei ein einstellbarer Parameter. Zu Beginn werden diese für die Initialisierung als Kodebuch festgesetzt. Die zufällige Initialisierung kann zu einer ungewollten Abweichung im Endergebnis des Lloyd Algorithmus führen (vgl. [Fin14] Kapitel 4). Eine Iteration des Lloyd Algorithmus geht dabei wie folgt vor :

- **Nächste Nachbar Bedingung:** Die Vektoren werden anhand des aktuellen Kodebuches C quantisiert. Das heißt jeder Vektor wird dem Zentroiden $c_i \in C$ zugeordnet, der die kleinste Distanz zu dem Vektor besitzt. Für das Erstellen des visuellen Vokabulars für die Bag of Features Repräsentation wird die euklidische Distanz genutzt.
- **Zentroid Bedingung:** Nachdem die Vektoren in die k Cluster eingeteilt wurden, wird das Kodebuch aktualisiert. Die neuen Zentroiden werden so gewählt, dass die Distanzen zu ihren Clustervektoren minimiert wird. Im Fall des Lloyd Algorithmus auf Basis der euklidischen Distanz werden diese durch die Bildung des Mittelwertvektors eines Clusters berechnet.

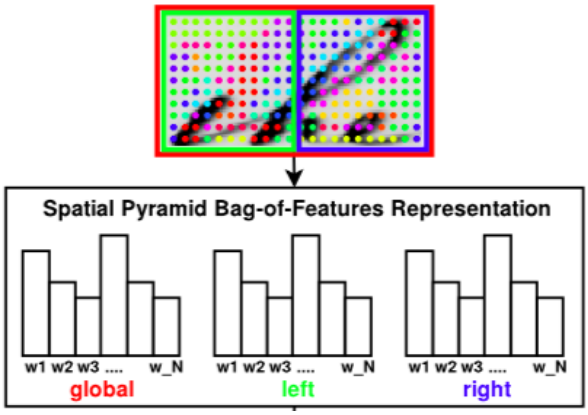



Abbildung 2.3.2: Spatial Pyramid mit einer 1x1 | 1x2 Auflösung: Bunte Punkte entsprechen den quantisierten Deskriptoren. Das Wort wird in zwei Hälften geteilt und zusätzlich zur globalen Bag of Features (roter Rahmen) wird für beide Teilregionen (grüner und blaue Rahmen) lokal ein Histogramm berechnet. Die Spatial Pyramid entsteht durch das Konkatenieren der Histogramme. Bild entnommen aus [RF16].

- **Terminierung:** Der Algorithmus bricht ab, wenn die Abbruchbedingung erfüllt ist. Dies kann gesteuert werden über die absolute Anzahl an Iterationen oder falls der relative Quantisierungsfehler konvergiert. Der Quantisierungsfehler ist die Summe der Distanzen zwischen den Vektoren und ihren zugewiesen Zentroiden. Wenn die Abbruchbedingung nicht erfüllt ist, wird eine weitere Iteration des Algorithmus durchgeführt.

Das generierte Kodebuch kann dann zur Quantisierung von SIFT Deskriptoren genutzt werden, indem zu einem Deskriptor der nächste Nachbar aus dem Kodebuch bestimmt wird.

### 2.4.2 *Spherical Lloyd*

Ein separates Clusteringverfahren wird für das in Abschnitt 1.2 beschriebene Gruppieren von Wortabbildern verwendet. Aufgrund der hochdimensionalen Spatial Pyramid Repräsentationen von Wortabbildern ist der Lloyd Algorithmus auf Basis der euklidischen Distanz nicht mehr geeignet. Es ist also ein Clustering Algorithmus für hochdimensionale Daten nötig. Ein Algorithmus hierfür ist der „Spherical k-Means“ [DM01, BDGS05, HFKB12]. In der Literatur wird der Lloyd Algorithmus fälschlicher-

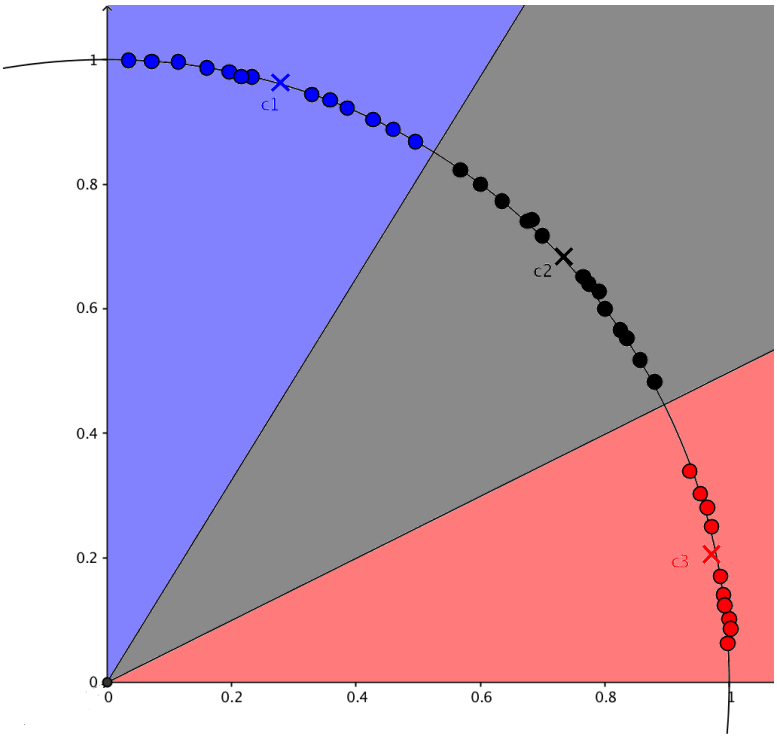



Abbildung 2.4.1: Die Zentroiden definieren die Cluster 1,2,3 (blau, schwarz, rot). Zu beachten ist in dieser Darstellung, dass alle Vektoren auf dem Einheitskreis des 1. Quadranten liegen, aufgrund der L2-Normalisierung und der ausschließlich positiven Eintrage der BoF Vektoren.

weise auch k-Means genannt. Genau genommen wird also ein „Spherical Lloyd“ Algorithmus benutzt, eine Variante des in Unterabschnitt 2.4.1 beschrieben Lloyd Algorithmus. Der Spherical Lloyd Algorithmus verwendet die Kosinus-Ähnlichkeit, ein winkelbasiertes Maß, als Distanzfunktion. Die Vektoren müssen vorher L2 normalisiert werden, sodass alle Vektoren die Länge 1 besitzen. Dadurch wird der Fokus auf die Richtung der Vektoren gelegt und die Länge der Vektoren ist dabei nicht mehr ausschlaggebend [DM01, BDGS05, HFKB12]. Die Kosinus-Ähnlichkeit zweier Vektoren berechnet sich wie folgt:

$$\cos(u,v) = \frac{\langle u,v\rangle}{\|u\| \cdot \|v\|} \qquad \forall u_i, v_i \in u,v : u_i, v_i >= 0, \quad \cos(u,v) \in [0,1] \quad (2.4.1)$$

Dabei bezeichnet $\langle u, v \rangle$ das Skalarprodukt zweier Vektoren $u, v$ und $\|u\|$ die L2-Normalisierung eines Vektors $u$. Da bei einer BoF Darstellung nur absolute Häufigkeiten von visuellen Wörtern betrachtet werden, können für die Kosinus-Ähnlichkeit nur Werte zwischen $0 - 1$ entstehen. Je größer der Wert für die Kosinus-Ähnlichkeit, desto ähnlicher sind sich zwei Vektoren. Daraus lässt sich dann die Kosinus Distanz ableiten mit $d_{cos}(u, v) = 1 - cos(u, v)$.

Bei einer Iteration des Algorithmus wird jeder Vektor dem nächsten Zentroiden des aktuellen Kodebuches zugewiesen (Nächste Nachbarbedingung). Dafür wird jeder Vektor $x$ der Eingabedatenmenge $X$ dem Zentroiden $c_i$ zugeteilt, der die Kosinus Distanz zu ihm minimiert. Nachdem die Vektoren in ihre Cluster eingeteilt wurden, wird das Kodebuch neu bestimmt. Die neuen Zentroiden entsprechen den Mittelwerten der Clustervektoren (Zentroidbedingung) [DM01, BDGS05, HFKB12].

Abbildung 2.4.1 visualisiert beispielhaft ein Ergebnis des Spherical Lloyd Algorithmus. Es wurden drei Cluster erzeugt (blau, schwarz, rot) mit ihren jeweiligen Zentroiden (Kreuze). Die ursprünglichen Datenpunkte befinden sich in den eingefärbten Bereichen. In der Abbildung sind die Vektoren nach ihrer L2-Normalisierung dargestellt und liegen somit auf dem Einheitskreis. Jeder Datenpunkt wird dem Zentroiden zugeordnet, der die kleinste Kosinus-Distanz zu ihm besitzt.

# 3 VERWANDTE ARBEITEN

Im vorherigen Kapitel wurden die wesentlichen Grundlagen für die segmentierungsfreie Indizierung erläutert. Das folgende Kapitel befasst sich mit den verwandten Arbeiten, die im Kontext der segmentierungsfreien Indizierung relevant sind. Bevor im nächsten Kapitel die vorgeschlagene Methodik präsentiert wird, ist das Ziel des aktuellen Kapitels, Verfahren mit einem starken Bezug sowohl zur segmentierungsfreien Indizierung als auch zur Methodik dieser Arbeit zu diskutieren.

Zu Beginn werden in Abschnitt 3.1 die relevanten fachlichen Schwerpunkte eingeführt. Die Textdetektion ist ein wichtiger Bestandteil (vgl. Kapitel 1) der segmentierungsfreien Indizierung und wird in Unterabschnitt 3.1.1 behandelt. Ein weiteres Forschungsgebiet ist das Word Spotting, dessen Ziel es ist visuell ähnliche Wörter zu einer Benutzeranfrage zu finden. Es handelt sich hierbei also um ein Unterproblem der Indizierung. Hierauf aufbauend werden die Herausforderungen der Indizierung in Zusammenhang mit dem Word Spotting gebracht.

In Abschnitt 3.2 werden Arbeiten vorgestellt, die einen direkten Bezug zur verwendeten Methodik und der segmentierungsfreien Indizierung haben. Da der MSER Detektor (siehe Abschnitt 2.1) hier Grundlage der Textdetektion ist, werden Vorgehen zur Textdetektion auf Basis von MSER vorgestellt. Danach werden segmentierungsbasierte Verfahren zur Indizierung und automatischen Transkription erläutert, deren Ergebnisse direkt mit der vorgeschlagenen Methodik vergleichbar sind. Anschließend werden einige Arbeiten zum segmentierungsfreien Word Spotting aufgeführt, da auch diese sich mit der Ähnlichkeitsbewertung von Wortabbildern im segmentierungsfreien Fall beschäftigen. Besonderer Fokus liegt dabei auf Patch-basierten Ansätzen [RRF13, RRLF14, RATL15, RATL11] und der Generierung von Worthypothesen durch Kombinieren von Textregionen [RSP09]. Zuletzt werden in der Schlussfolgerung (Abschnitt 3.3) dieses Kapitels die diskutierten Arbeiten mit der Methodik der vorliegenden Arbeit direkt in Bezug gesetzt.

## 3.1 FACHLICHE SCHWERPUNKTE

Bevor im Detail spezifische Arbeiten erläutert werden, sollen die wichtigen verschiedenen Fortschritte in den drei relevanten Forschungsfelder kurz vorgestellt werden.

Im segmentierungsfreien Fall ist keine Segmentierung auf Wortebene bekannt. Deswegen ist die Textdetektion nötig, um die Wortabbilder eines Dokumentes zu finden. Darauf wird das Konzept des Word Spotting erläutert und in Bezug mit der Indizierung gesetzt.

### 3.1.1 *Textdetektion*

Der Bereich der Textdetektion beschäftigt sich mit der Lokalisierung von Wörtern in Bildern. Klar voneinander abzugrenzen sind dabei die unterschiedlichen Anwendungsgebiete. Im Bereich der Analyse von modernen Dokumenten existieren Verfahren, wie die OCR (vgl. Kapitel 1), die hierfür adäquate Ergebnisse liefern. Maschinengeschriebene Dokumente besitzen üblicherweise einen einheitlichen Schriftsatz. Zudem sind Wörter eindeutig voneinander trennbar, da in diesen Überlappungen selten sind. Hintergründe sind meist einheitliche weiße Flächen, womit eine simple Unterteilung in Hintergrund und Text möglich ist. Diese Eigenschaften vereinfachen die Textdetektion in modernen Dokumenten.

Im Gegensatz hierzu gilt die Segmentierung von handgeschriebenen historischen Dokumenten weiterhin ungelöst, da die oben genannten Bedingungen nicht übertragbar sind. Handschrift in historischen Dokumenten ist oft nicht klar strukturiert, die Schriftgröße kann unteranderem nicht einheitlich sein. Genauso können Abstände zwischen Wörtern stark variieren. Zusätzlich besteht die Möglichkeit, dass Wörter verschiedener Zeilen sich überlappen. Das Alter der Dokumente kann zudem einen großen Einfluss auf den Qualitätszustand des Papiers und die Schrift haben.

Ein weiteres Anwendungsgebiet der Textdetektion ist die innerhalb von natürlichen Szenenbilder, in denen sich ein Großteil der aktuellen Arbeiten einreihen lassen [WBB11, CCC+11, CTS+11, NM11a, QM16]. Auch hier sind die Annahmen über moderne Dokumente nicht gegeben. Problematisch ist die große Variation in den Hintergründen. Es gilt zwischen Alltagsobjekten wie z. B. Menschen und Autos, die in diesem Fall dem Hintergrund zuzuordnen sind und relevanten Textstellen zu unterscheiden. Die erschwerten Rahmenbedingungen in diesem Szenario motivieren dazu, ähnliche Verfahren für den Fall der historischen Dokumente anzuwenden. Zunächst werden jedoch Ansätze vorgestellt, die im Zusammenhang von realen Szenenbildern entwickelt wurden.

#### *Textdetektion in natürlichen Szenen*

Wang et al. benutzen zur Textdetektion in ihrer Arbeit [WBB11] ein Verfahren auf Basis eines gleitendes Fensters. Das Fenster wird für die Detektion von Buchstaben

genutzt. Dafür wird ein Fenster über das Bild geschoben, um alle Positionen abzudecken. Dabei ist es nötig, diesen Prozess mit verschiedenen Fenstergrößen durchzuführen, um Zeichen unterschiedlicher Abmessungen lokalisieren zu können. Für die Bildbereiche, die jeweils durch ein Fenster eingegrenzt sind, gilt es zu entscheiden, ob es sich bei diesen um Zeichen oder Hintergrund handeln.

Die Autoren nutzen hierfür ein Multiklassen-Klassifikator mit 62 Klassen (52 Buchstaben und 10 Ziffern), den Random Fern, eine Variante des Random Forest Klassifikators [WBB11, BZMn07]. Aus den klassifizierten Zeichen werden die wahrscheinlichsten Aneinanderreihungen zu Wörtern berechnet [WBB11].

Ein gleitendes Fenster wird auch von Coates et al. in [CCC+11] zur Textdetektion benutzt, mit dessen Hilfe mögliche Textregionen zur Klassifizierung in Text- oder Hintergrundstellen ermittelt werden. Zur Klassifizierung von möglichen Textbereichen wird eine lineare Support Vector Machine (SVM) trainiert. Eine SVM ist ein diskriminativer Klassifikator, der in einer Trainingsphase eine Hyperebene berechnet. Dafür sind Trainingsdaten nötig, für die eine korrekte Klassenzuordnung vorausgesetzt wird. Das Ziel der Hyperebene ist es, die Datenpunkte verschiedener Klassen bestmöglich voneinander zu trennen. Durch die Lage eines Vektors im Merkmalsraum wird anhand der Hyperebene entschieden, ob es sich bei diesen um Text bzw. Hintergrund handelt (Näheres zu SVM in [Bur98]).

*Textdetektion in historischen Dokumenten*

Der folgende Abschnitt behandelt Arbeiten, welche die Textdetektion in historischen Dokumenten betrachten. Anstelle eines gleitenden Festers wird in [KWD14] eine heuristische Wortdetektion vorgeschlagen. Zusammenhangskomponenten (vgl. Unterabschnitt 2.1.2) werden anhand eines Schwellwertverfahrens berechnet. Mit den Komponenten werden mithilfe heuristischer Regeln Worthypothesen erzeugt. Komponenten, die entlang der x- oder y-Achse zu weit auseinander liegen, werden nicht zusammengefasst. Eine weitere Regel bezieht sich auf die Abmessungen von Hypothesen. Zu große, breite oder hohe Hypothesen werden verworfen.

Manmatha und Rothfeder präsentieren in [MR05] ein skalenbasiertes Segmentierungverfahren. In einem Vorverarbeitungsschritt werden automatisch unerwünschte Ränder, die beim Scannen der Dokumente entstanden sind und horizontale Linien, die vom Schreiber als Unterstreichungen von wichtigen Zeilen eingefügt worden sind, entfernt (siehe Abbildung 3.1.1).

Auf Basis der horizontalen Projektionsprofile (siehe Unterabschnitt 3.2.2) wird im nächsten Schritt das Dokument in Zeilen unterteilt. Für horizontale Projektionsprofile werden die Grauwertintensitäten reihenweise addiert. Obere und untere Zeilen-

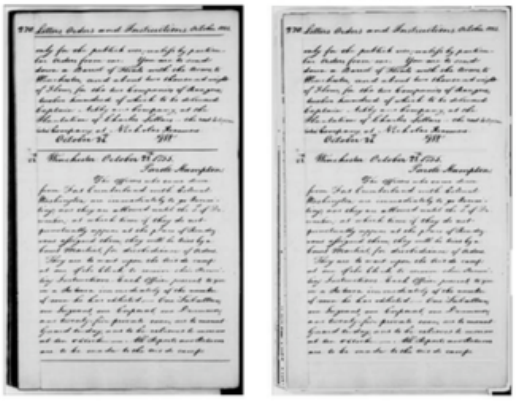


Abbildung 3.1.1: Im linken Teil sind bei der Digitalisierung Artefakte entstanden, wie an den Rändern zu sehen ist. In einem Vorverarbeitungsschritt werden diese und Unterstreichungen entfernt. Das rechte Bild zeigt das Dokument nach der Vorverarbeitung. Bild entnommen aus [MR05].

begrenzungen zeichnen sich dadurch aus, dass diese größtenteils aus weißen Hintergrundpixeln bestehen. Es werden also lokale Maxima gesucht, die mit Zeilen korrespondieren. Abbildung 3.1.2 zeigt den Verlauf der Zeilensegmentierung. Die lokalen Maxima des Projektionsprofil korrespondieren mit Zeilenbegrenzungen und werden als Zeileneinteilung gewählt.

Mithilfe der Zeilensegmentierung werden zeilenweise Blobs (Unterabschnitt 2.1.2) generiert. Zur Generierung der Blobs wird ein neuer Filter definiert. Es handelt sich hierbei um eine Variante des Laplacian of Gaussian (LOG) Filters. Ein LOG Filter approximiert die Ableitungen zweiter Ordnung der Gaußfunktion in x- und y-Richtung und kann als Kantendetektor verwendet werden (vgl. [Jä12] Seite 390-391). Im diskreten Fall wird er angenähert durch Faltungen mit einem Gauß- und anschließend einem Laplace-Filter. Wörter haben häufig ein Seitenverhältnis größer als 1 und sind somit breiter als hoch. Der neu definierte Filter ist deswegen anisotrop statt isotrop, mit einer größeren Ausdehnung in x- als in y-Richtung [MR05].

In Abbildung 3.1.3 ist der Einfluss von verschiedenen Skalierungen auf die Blobgenerierung visualisiert. Kleine Skalierungen erzeugen kleine Blobs, die einzelnen Zeichen entsprechen, während große Skalierungen zum Verschmelzen von ganzen Wörtern führen. Es gilt einen passenden Wert für die Ausbreitung in x- und y-Richtung zu finden, sodass Blobs mit holistischen Wörtern korrespondieren. Um dem Problem der Optimierung von zwei Parametern zu entgehen, wird das Verhältnis von x und y konstant festgesetzt. Somit reicht es nach dem Parameter für y zu suchen, da sich

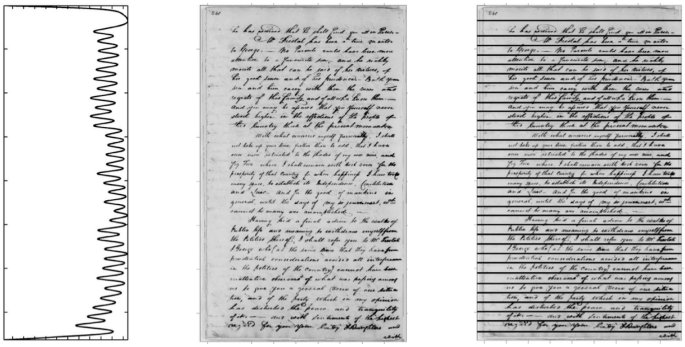


Abbildung 3.1.2: Das erste Bild zeigt das horizontale Projektionsprofil des mittleren Dokumentes. Lokale Maxima sind Zeilen, mit wenig Schriftpixeln und korrespondieren mit Zeileneinteilung. Bild entnommen aus [MR05].

aus diesem dann, durch das konstante Verhältnis, auch x bestimmen lässt. Der Wert nach dem optimiert wird, ist die Fläche A, welche die Blobs innerhalb einer Zeile ausmachen. Dabei wird, wie in Abbildung 3.1.3 zu sehen die Menge der weißen Pixel betrachtet und maximiert. Die Abbildung verdeutlicht, dass die Zusammenhangskomponenten eher Wörtern entsprechen, wenn die Fläche A groß ist (d)-(e).

Für die Bestimmung der Ausbreitung in y-Richtung sind mehrere Verfahren präsentiert worden. Der erste Ansatz sucht innerhalb eines größeren Intervalls nach dem optimalen Wert. Die Suche innerhalb des Intervalls ist rechenintensiv, da viele Werte für y getestet werden müssen. Das zweite Vorgehen approximiert den Wert anhand der durchschnittlichen Zeilenhöhe über den gesamten Korpus. Der letzte Ansatz betrachtet die durchschnittliche Zeilenhöhe innerhalb des aktuellen Dokumentes und nicht über alle des Korpus.

Manmatha und Rothfeder hatten als Ziel eine optimale Segmentierung der Wörter zu erzeugen. Dies ist, wie zu Beginn des Abschnittes gezeigt, im Kontext der historischen Dokumente keineswegs trivial. Im Gegensatz zu diesem Verfahren wird hier eine weichere Voraussetzung angenommen und in einem hypothesenbasierten Verfahren verschiedene Varianten für mögliche Wortabbilder erzeugt.

### 3.1.2 *Word Spotting*

Das Gebiet des Word Spotting für historische Dokumente befasst sich mit der Ähnlichkeitsbewertung von Wortabbildern. Ursprünglich in [MHR96] zur Indizierung von historischen Dokumenten vorgeschlagen, wird der Begriff Word Spotting heutzutage


Abbildung 3.1.3: Blobgenerierung mithilfe des definierten Filters. Bei kleinen Skalierungen entsprechen Blobs einzelnen Zeichen (b). Größere Skalierungen erzeugen Blobs, die eher mit Wörtern korrespondieren (c)-(e). Sind die Werte zu groß gewählt kann das Ergebnis der Blobgenerierung fehlschlagen (f)-(g). Es gibt eine Skalierung, sodass die Blobs annähernd mit Wörtern korrespondieren. Bild entnommen aus [MR05].

für Verfahren zum Retrieval verwendet. Aus diesem Grund gilt es im Folgenden die Aufgabenbereiche des Word Spotting klar von einander abzugrenzen. Zunächst wird jedoch Word Spotting zum Retrieval erläutert.

Das automatische Transkribieren von historischen Dokumenten ist, wie schon in Kapitel 1 gezeigt, ein schwieriger Prozess, bei dem es aktuell immer noch zu hohen Fehlerraten kommen kann. Ein populäres Vorgehen der Dokumentenanalyse diesem Problem zu entgehen, ist das sogenannte Query-by-Example (QbE) Word Spotting. Hierbei wird auf das Erkennen der Wörter verzichtet. Stattdessen wird vom Nutzer eine Beispielanfrage (Query) gestellt, indem zu der Anfrage die visuell ähnlichsten Wörter aus der Sammlung gesucht werden, kann eine Transkription und das damit verbundene Erkennen der Wörter vermieden werden. Ähnlich zu einer Textsuche in modernen Texteditoren sollen so alle Vorkommen der Anfrage dem User angezeigt werden. Somit ist das Word Spotting dem Gebiet des Information Retrieval zuzuordnen. Ansätze zum Word Spotting lassen sich in zwei Kategorien einteilen:

Der segmentierungsbasierte Fall setzt voraus, dass im Vorfeld eine Wortsegmentierung stattgefunden hat und fokussiert sich auf den Vergleich von Wortabbildern. Das Problem im segmentierungsbasierten Fall ist, dass von einer fehlerfreien Segmentierung ausgegangen wird und wie schon in Unterabschnitt 3.1.1 erwähnt, eine optimale Segmentierung keineswegs trivial ist. Somit hat die Qualität der Segmentierung einen starken Einfluss auf die Ergebnisse des Word Spotting [DNLP16, RATL15, RATL11]. Zu einer Anfrage werden dann die segmentierten Wortabbilder verglichen und die ähnlichsten Kandidaten in der sogenannten Retrieval Liste zurückgegeben.

Im segmentierungsfreien Word Spotting wird keine explizite Wortsegmentierung durchgeführt. Dokumente werden als ganzes betrachtet und das Ziel ist es, die zur Anfrage relevanten Bildbereiche zu extrahieren. Somit werden mögliche Segmentierungsfehler umgegangen, die bei der Ähnlichkeitsbewertung zwischen Anfrage und Wortkandidaten zu Fehlern führen können. Erfolgreiche Ansätze im segmentierungsfreien Fall arbeiten Patch-basiert. Dafür wird das zu analysierende Dokument in regelmäßige Bildbereiche unterteilt [RATL15, RRF13, RRLF14, RATL11]. Unterabschnitt 3.2.5 wird einige relevante segmentierungsfreie Ansätze zum Word Spotting diskutieren.

Das Prinzip des Finden von visuell ähnlichen Wörtern ist auch für die Indizierung von großer Bedeutung. Somit lässt sich Word Spotting als Unterproblem der eigentlichen Indizierung auffassen. Im nächsten Abschnitt werden die Aspekte des Word Spotting zur Indizierung behandelt.

### 3.1.3 *Word Spotting: Indizierung*

In der Literatur hat sich der Begriff des Word Spotting für die in Unterabschnitt 3.1.2 beschriebene Retrieval Aufgabe etabliert. Die wesentlichen Parallelen und Unterschiede zur Indizierung sollen in diesem Abschnitt diskutiert werden. Analog zum Word Spotting ist für die Indizierung die Wahl der Merkmalsdarstellung von Wortabbildern essentiell. Zusätzlich ist ein geeignetes Maß für die Beurteilung der Ähnlichkeit von Repräsentationen nötig, sodass die Bewertung zweier Wörter besser ausfällt, wenn diese visuell ähnlich sind.

Im Gegensatz zum Retrieval, in dem vom Nutzer eine Anfrage gestellt wird und alle Wortkandidaten mit der Anfrage verglichen werden, um die ähnlichsten Wörter in der Retrieval Liste zurückzugeben, werden für die Generierung eines Indexes keine expliziten Anfragen definiert. Stattdessen ist das Ziel, die Menge der Wortabbilder in Gruppen zu unterteilen, sodass sich innerhalb jeder Gruppe nur Vorkommen des gleichen Wortes befinden. Ein Verfahren hierfür ist das Clustering (siehe Abschnitt 2.4).

Das Konzept der Ähnlichkeitsbewertung des Word Spotting findet sich im Clustering wieder. Als Pendant zur Nutzeranfrage lässt sich die Zentroidbedingung (vgl. Abschnitt 2.4) auffassen. Mit den Zentroiden als Modell, werden die visuell ähnlichsten Wortabbilder gesucht. Die Bewertung der Ähnlichkeit zwischen Wortabbildern spiegelt sich in der nächsten Nachbarbedingung wider. Jedes Wortabbild wird im Optimalfall dem Cluster zugeordnet, zu dessen Wortabbildern die größte visuelle Übereinstimmung herrscht.

Ein anderes Verfahren zur Clusteranalyse ist das agglomeratives Prinzip. Agglomeratives Clustering beginnt mit je einem Cluster pro Datum (Wortabbild). Schrittweise werden die Cluster zusammengelegt, sodass immer größere Cluster entstehen. Pro Schritt werden die beiden Cluster mit der größten sogenannten Inter-Cluster Ähnlichkeit ermittelt und vereint. Dabei beschreibt das Inter-Cluster Maß die Ähnlichkeit zwischen zwei Clustern (vgl. [HTF01] Seite 520 ff.). In [RM06] werden unter anderem agglomerative Verfahren für die Indizierung genutzt.

## 3.2 SPEZIELL RELEVANTE ARBEITEN

Der Fokus des folgenden Kapitels wird auf Arbeiten liegen, die besonders zu der verwendeten Methodik relevant sind. Dabei werden insbesondere in Unterabschnitt 3.2.1 Ansätze zur Textdetektion auf Basis des MSER Detektors behandelt, da dieser Bestandteil der vorgeschlagenen Methodik ist. Anschließend wird die Arbeit [RM06] vorgestellt, welches die grundlegende Idee des Clusterings zur Generierung eines Index hier motiviert. In Unterabschnitt 3.2.3 und 3.2.4 werden zwei Verfahren zur automatischen Transkription von historischen Dokumenten behandelt. Da auch mit der Indizierung eine Transkription erreichbar ist (vgl. Abschnitt 1.2), haben diese einen starken Bezug zur vorliegenden Arbeit. Unterabschnitt 3.2.5 beschreibt Ansätze zum segmentierungsfreien Word Spotting.

### 3.2.1 *Textdetektion auf Basis des MSER Detektor*

Der MSER Detektor eignet sich besonders zur Textdetektion (vgl. Abbildung 2.1.1 und [NM11b, CTS+11, NM11a, TD15, NM11b]) und wird deswegen hier zur Lokalisierung von Textregionen verwendet. Aus diesem Grund werden im Folgenden Arbeiten zur Textdetektion auf Basis des MSER Detektor vorgestellt.

Matas und Neumann stellen in [NM11a] ein Verfahren zur Textdetektion innerhalb natürlicher Szenen vor. Der MSER Detektor wird angewendet, um die einzelnen Buchstaben im Bild zu extrahieren. Dabei wird der MSER Detektor nicht nur auf das

Graustufenbild angewendet, sondern im Weiteren das MSER Verfahren auch auf den einzelnen Rot, Grün und Blau Kanälen durchgeführt. Die Resultate des MSER Detektors können auch Fehldetektionen beinhalten, also Bereiche, die nur Hintergrund umfassen. Für die Entscheidung, ob eine MSER Region einen Buchstaben enthält, wird eine SVM verwendet, welche Regionen anhand geometrischer Eigenschaften wie zum Beispiel Höhen- und Seitenverhältnis etc. als Hintergrund bzw. Text klassifiziert. Für das Training der SVM wurden MSER Regionen auf Bildern der Internet Community Flickr[1] berechnet. Die Trainingsmenge wurde dann manuell in Buchstaben und Hintergrund unterteilt. Die als Buchstaben klassifizierten Regionen werden zu Wörtern gruppiert. Dafür werden aus den gefundenen Buchstaben, Sequenzen durch Kombinieren der einzelnen Buchstaben erstellt. Diese erzeugen einen potentiellen Suchraum von Worthypothesen, der exponentiell wächst. Um diesem Problem zu entgehen wird der Suchraum verkleinert, indem nur Buchstabensequenzen, die als zulässige Teilwörter identifiziert werden, zur weiteren Sequenzgenerierung betrachtet werden. Sequenzen, die keinen Teilwörtern entsprechen werden aussortiert, womit der exponentielle Anstieg an erstellten Wortkandidaten verhindert wird. Damit dies möglich ist, muss für eine Sequenz entschieden werden, ob sie einer Textregion entspricht. Für zweielementige Sequenzen werden paarweise Bedingungen überprüft. Dafür werden Werte wie zum Beispiel die horizontale Distanz und der Höhenunterschied zwischen den beiden Regionen berechnet und Sequenzen verworfen, wenn diese Werte zu groß sind.

Unter der Annahme, dass Wörter sich sowohl durch maximal zwei obere als auch zwei untere Linien eingrenzen lassen, werden für dreielementige Sequenzen Wortlinien erzeugt. In Abbildung 3.2.1 ist dies visuell dargestellt. Anhand von verschiedenen Heuristiken wird geprüft, ob diese konsistent sind. Insbesondere werden dafür die vertikalen Distanzen der Wortlinien untereinander betrachtet (vgl. Abbildung 3.2.1).

Es wurde gezeigt, dass für Sequenzen, die länger als drei sind, kein Gewinn von Genauigkeit bei der Generierung von Wortlinien erzielt wird [NM11a]. Mit dieser Annahme werden im weiteren Sequenzen größer als drei ausgewertet, indem alle dreielementigen Teilsequenzen auf Konsistenz überprüft werden. Die Menge der lokalisierten Wörter entspricht den Sequenzen, die als Textregion identifiziert wurden und nicht durch weitere Kombinationen erweiterbar sind.

Eine weitere Arbeit auf Basis des MSER Detektors ist von Tabassum und Dhondse präsentiert worden [TD15]. Um die Anfälligkeit des MSER Detektors durch unscharfe Bilder (vgl. [MTS+05]) zu umgehen, wird dieser mit dem Canny Edge Algorithmus, welcher zur Kantendetektion genutzt wird, kombiniert (vgl. [SC+12]). Dafür werden

1 Social Media Seite: Zu finden auf https://www.flickr.com

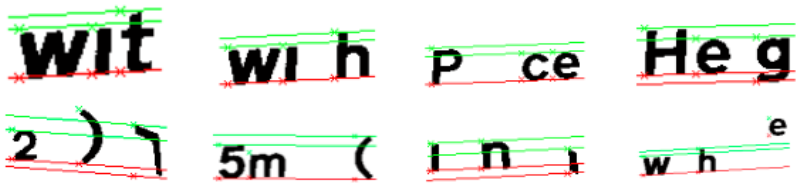


Abbildung 3.2.1: Oberen (grün) und unteren (rot) Linien sind hier für jeweils dreiementige Sequenzen erzeugt worden. Die Sequenzen der oberen Reihe wurden akzeptiert und die ganze untere Reihe abgelehnt. Bild entnommen aus [NM11a].

die Randpixel der MSER Regionen neu definiert und von den gefundenen Kantenpixel des Canny Edge Algorithmus ersetzt. Somit werden alle Pixel, die vermutlich nicht zur Kante und damit nicht zum Schriftverlauf gehören aus der MSER Region entfernt. Regionen, die gewisse geometrische Bedingungen nicht erfüllen, werden verworfen (Seitenverhältnis etc.). Zusätzlich hierzu wird die Stroke Width Transform [EOW10] genutzt, welche zu jedem Pixel die Breite des zu dem Pixel wahrscheinlichsten Schriftverlauf berechnet. Somit wird für jeden Pixel die zugehörige Strichstärke ermittelt. Regionen, die ähnliche Strichstärken besitzen und nicht zu weit auseinander liegen, werden dann zu Wörtern gruppiert.

Auch im Bereich der Neuronalen Netze existieren Arbeiten, die Gebrauch vom MSER Detektor machen. Qin und Manduchi klassifizieren MSER Regionen mithilfe eines Convolutional Neural Network (CNN) [QM16]. Ein CNN besteht aus mehreren Schichten, die über Faltungen an den verschiedenen Schichten Merkmalsvektoren erzeugen. Am Ende der Architektur liegt im Fully Connected Layer ein Klassifikator, der den Vektor in Text bzw. Hintergrund einteilt. Die als Text klassifizierten Regionen werden im letzten Schritt dann zu Wörtern kombiniert.

### 3.2.2 *Indizierung von historischen Dokumenten*

Rath und Manmatha haben in [RM06] eine Methodik präsentiert, um historische Dokumente zu indizieren. Diese Arbeit dient als grundlegende Motivation für die wesentliche Idee der vorliegenden Arbeit. Das Prinzip des Clustering von Wortabbildern wird auch hier verwendet. Das Verfahren ist für den segmentierungsbasierten Kontext vorgestellt worden und setzt einen vorherigen Segmentierungsschritt voraus. Zu Beginn dieses Abschnittes wird eine weitere alternative Merkmalsdarstellung von Wortabbildern erläutert.

*Alternative Wortrepräsentation zur Bag of Features*

Vor dem Beginn der Merkmalsextraktion werden zunächst die Wörter einem Vorverarbeitungsschritt unterzogen. Um das Problem der in Kapitel 1 beschriebenen unerwünschten Variabilität im visuellen Vorkommen gleicher Wörter zu eliminieren, werden die Wortabbilder normalisiert, indem eventuelle Zeilenneigungen (Skew) und Schriftneigungen (Slant) entfernt werden. Nach der Normalisierung werden für die Wortabbilder Merkmalsrepräsentationen berechnet. Dafür werden einerseits Projektionsprofile berechnet, welche die Verteilung der Schriftpixel entlang der Spalten modellieren sollen. Diese werden bestimmt, indem die Grauwertintensitäten spaltenweise addiert werden. Zusätzlich werden Wort Profile berechnet, welche die äußeren geometrischen Schriftverläufe repräsentieren. Mithilfe eines einfachen Schwellwertverfahrens wird der Bilddausschnitt in Schrift- und Hintergrundpixel unterteilt. Obere und untere Wortprofile lassen sich definieren, indem für jede Spalte des Bildes der oberste bzw. unterste Schriftpixel zugewiesen wird. Zur Modellierung von inneren Schriftverläufen wird ein weiteres Merkmal (Background/Ink Transition) berechnet, welches die Wechsel von Schrift- und Hintergrundpixel für jede Spalte erfasst [RM06, RM03].

Die beschriebenen Merkmalsrepräsentationen betrachten Wortabbilder spaltenweise. Durch die sequentielle Darstellung haben die Merkmalsvektoren der Wortabbilder unterschiedliche Dimensionen. Im Vergleich zu kürzeren Wörtern bestehen längere Wortabbilder aus mehr Spalten und es entstehen Vektoren unterschiedlicher Dimensionen. Aus diesem Grund wird der Dynamic Time Warping (DTW) Algorithmus verwendet, um den Grad der Übereinstimmung zweier Wortabbilder zu bewerten. Der DTW Algorithmus basiert auf dynamischer Programmierung und bestimmt die Kosten, um ein Signal auf ein anderes abzubilden (für eine ausführliche Beschreibung des DTW sei auf [RM06] verwiesen).

Eine weitere alternative Repräsentation wird verwendet, um einen einheitlichen Merkmalsraum zu generieren, womit die rechenintensive DTW vermeidbar ist. Auf den Profilen wird dafür eine diskrete Fourier Transformation (DFT) angewendet und die ersten $k$ Koeffizienten dieser genommen, um Merkmalsvektoren gleicher Länge zu erstellen. Diese können dann mit herkömmlichen Distanzmaßen verglichen werden.

*Word Spotting zur Indizierung*

Das Gruppieren von Wörtern wird mithilfe eines Clustering Algorithmus durchgeführt. Für solch einen Algorithmus ist es notwendig, eine Aussage über die Anzahl

der Cluster zu treffen. Die Autoren schlagen an dieser Stelle Heap's Gesetz vor, eine empirisch belegte Regel aus dem Bereich der Linguistik, welches die Anzahl an unterschiedlichen Wörtern einer Sammlung von Wörtern schätzt [RM06]. Sei $N$ die Anzahl aller Wörter des Korpus, so lässt sich die Anzahl an einmaligen Wörtern $V(N)$ approximieren mit: $V(N) = K \cdot n^{\beta}$. Dabei sind $K$ und $\beta$ sprach- und kontextabhängige Parameter. Die Autoren nutzten das Nelder-Mead Optimierungsverfahren, um $K$ und $\beta$ zu schätzen (Näheres in [NM65]).

Für die Indizierung werden mehrere Clustering Algorithmen verglichen. Der in Unterabschnitt 2.4.1 beschriebene Lloyd Algorithmus wird verschiedenen agglomerativen Clustering Verfahren gegenübergestellt. Die verschiedenen Algorithmen werden auf zwei Weisen durchgeführt. Zum einen wird das Clustering mit der sequentiellen Darstellung von Wortabbildern und der DTW Distanz ausgeführt. Da die DTW sehr rechnenintensiv ist, werden die paarweisen DTW Distanzen vorher berechnet und in einer Matrix abgespeichert, damit diese nicht zur Laufzeit des Clustering berechnet werden müssen. Dies hat zur Folge, dass mit dieser Repräsentation der Lloyd Algorithmus nicht ausgewertet werden kann. Aus der Distanzmatrix sind die Merkmalsvektoren der Wörter nicht mehr rekonstruierbar, womit die Clusterzentroiden nicht berechnet werden können. Zusätzlich werden die Algorithmen mit der DFT Darstellung der Wortabbilder durchgeführt. Das beste Ergebnis wurde mit Ward's Linkage auf den DFT Koeffizienten der Profile erreicht. Ward's Linkage ist ein agglomerativer Algorithmus, der als Inter-Cluster Maß die Summe der quadratischen Abweichung betrachtet [ML11].

Für die Generierung eines Index sind nicht alle Cluster relevant. Es gilt die Cluster zu identifizieren, die eine wichtige Bedeutung für den Text haben. Hierfür wird ein weiteres empirisches Gesetz der Sprachwissenschaft vorgeschlagen. Mit Hilfe des sogenannten Zipf'schen Gesetzes lassen sich Aussagen über die Frequenzen von Wortvorkommen formulieren [Pia14]. Das Zipf'sche Gesetz sagt aus, dass die Häufigkeit des Vorkommens sich antiproportional zur Relevanz eines Wortes verhält. Durch diese Modellierung können die relevanten Stichwörter des zu generierenden Index identifiziert werden. Mit hoher Wahrscheinlichkeit entsprechen Ausdrücke, die sehr häufig auftreten semantisch unbedeutenden Stoppwörter. Es ist also sinnvoll Cluster, die aus vielen Einträgen bestehen zu verwerfen. Zusätzlich wurde die These aufgestellt, dass Wörter nicht zentral für den vorliegenden Text sind, wenn diese nur selten auftreten. Diese Annahme ist an dieser Stelle fraglich, da vor allem im Kontext der historischen Dokumente z. B. Nennungen außergewöhnlicher Ereignisse selten auftauchen, aber diese dennoch von großer Relevanz sind.

### 3.2.3 *Handschrifterkennung in handgeschriebenen historische Dokumente mit HMMs*

Lavrenko et al. stellen in [LRM04] ein segmentierungsbasiertes Verfahren vor, mit dem Ziel historische Dokumente automatisch zu transkribieren. Im Gegensatz zum Query-by-Example Word Spotting ist es nötig, die Wörter der Dokumente zu erkennen. Eine Methode zur Handschrifterkennung stellt das Hidden Markov Modell (HMM) dar. Anzumerken ist, dass das HMM dort nicht auf die traditionelle Weise zur Handschrifterkennung verwendet wird. Bei der traditionellen Funktionsweise des HMMs zur Handschrifterkennung wird eine sequentielle Merkmalsrepräsentation für ein Wortabbild bestimmt. Dafür wird oft ein gleitendes Fenster benutzt. Innerhalb des Fensters wird ein Merkmalsvektor bestimmt. Durch das Bewegen des Fensters auf Zeilenebene entsteht eine sequentielle Merkmalsdarstellung für ein Wortabbild. Mit dieser beobachteten Sequenz wird das Modell bestimmt, welches am wahrscheinlichsten diese Sequenz erzeugt hat und anhand dessen das Wort klassifiziert (vgl. [Fin14] Kapitel 14).

Das HMM in [LRM04] geht stattdessen holistisch vor. Es werden keine Sequenzen von Merkmalsvektoren für ein Wort betrachtet, sondern Folgen holistischer Merkmalsdarstellungen von Wortabbildern. Für die holistische Merkmalsdarstellung eines Wortes werden triviale geometrische Eigenschaften benutzt (Höhe, Breite, Seitenverhältnis, etc.) und zusätzlich die DFT Koeffizienten der in Unterabschnitt 3.2.2 beschriebenen Profile.

Abbildung 3.2.2 visualisiert das verwendete HMM. Die Zustände des HMM werden als die Wörter der Dokumente gewählt. Diese Zustände sind nicht direkt beobachtbar, zu jedem Zeitpunkt wird aber im aktuellen Zustand ein beobachtbarer Merkmalsvektor erzeugt, der eine holistische Merkmalsrepräsentation dieses Wortes darstellt. Für die Bestimmung einer Transkription wird ein Dokument als Sequenz von Merkmalsvektoren seiner Wörter betrachtet. Mithilfe des Modells wird die wahrscheinlichste Reihenfolge der Wörter (Zustandsübergänge) ermittelt, welche die beobachtete Sequenz generiert hat, die dann für die Transkription übernommen wird. Ein Zustandsübergang modelliert somit das Folgen eines neuen Wortes auf das aktuelle Wort in einem Dokument. In einer Trainingsphase werden die Übergangswahrscheinlichkeiten zwischen Zuständen und die Wahrscheinlichkeiten für bestimmte Beobachtungen für jeden Zustand des HMM geschätzt. Dabei ist eine größere Menge an Trainingsdaten nötig. Für eine ausführliche Erklärung zum HMM sei hier auf [Fin14] verwiesen.

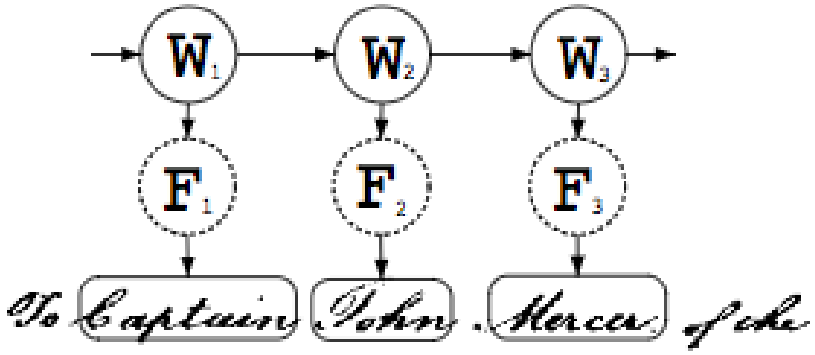



Abbildung 3.2.2: Die Zustände entsprechen den Wörtern des Trainingsvokabulars $(W_1, W_2, W_3)$. Zustandsübergänge sind als Pfeile zwischen den $W_i$ dargestellt. Sie beschreiben das Folgen eines neuen Wortes. In jedem Zustand wird eine Ausgabe generiert $(F_1, F_2, F_3)$. Zu einer beobachteten Sequenz von $F_i$ wird die Folge von Wörtern $W_i$ berechnet, die diese am wahrscheinlichsten erzeugt hat. Diese Reihenfolge von Wörtern wird als Transkription gewählt. Abbildung entnommen aus [LRM04].

### 3.2.4 *Handschrifterkennung in handgeschriebenen historischen Dokumenten mit Entscheidungsbäumen*

Howe et al. präsentieren in [HRM05] ein segmentierungsbasiertes Verfahren zur Generierung einer Transkription von Dokumenten. Auch hier ist das Ziel ein Modell zur Handschrifterkennung in historischen Dokumenten zu erstellen. Dafür wird jedes einzigartige Wort in einer eigenen Klasse dargestellt. Für ein zu erkennendes Wortabbild gilt es dann, die richtige Wortklasse zu identifizieren. Ein geeignetes Verfahren zur Klassifikation bei einer größeren Anzahl von Klassen sind Entscheidungsbäume (vgl. [SL91]). Abbildung 3.2.3 visualisiert schematisch die Klassifikation mit einem Entscheidungsbaum. Vom Startknoten aus werden an den besuchten Knoten Kriterien überprüft. Je nach Auswertung der Bedingungen wird an den Knoten verzweigt, bis ein Blattknoten erreicht wird. Die Blattknoten korrespondieren mit den einzelnen Wortklassen, somit ist der erreichte Blattknoten dann die Klassifikation des Wortabbildes.

Für die Klassifikation der Wortabbilder in die entsprechenden Wortklassen wird ein Ensemble von Entscheidungsbäumen mit dem AdaBoost Algorithmus trainiert.

Bäume können in der Theorie mit einer Klassifikationsgenauigkeit der Stichprobe von 100% trainiert werden. Dies führt in der Regel aber zu Modellen, die überan-

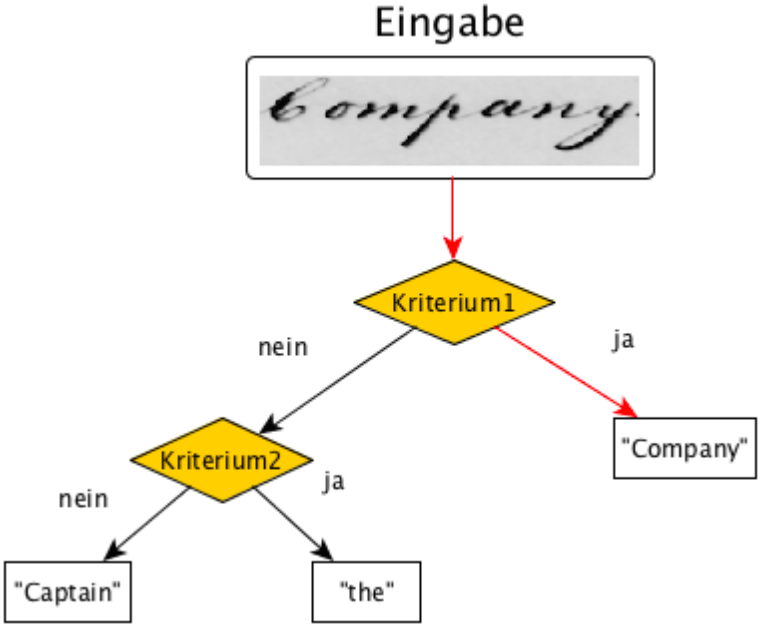



Abbildung 3.2.3: Beispielhafte Darstellung eines Entscheidungsbaum. Das Wortabbild "Company" soll klassifiziert werden. Die Blätter des Baumes sind als rechteckige Boxen dargestellt, Sie entsprechen den Wortklassen „Company“, „the“, „Captain“. An den besuchten Knoten wird verzweigt (rote Pfeile), bis der Blattknoten mit der Klasse "Company" erreicht wird. Diese ist dann die zugeteilte Transkription des Wortabbildes.

gepasst sind (Overfitting) [HRM05]. Überangepasste Modelle erreichen zwar auf der Trainingsmenge gute Klassifikationsergebnisse, können von dieser aber nicht weit genug abstrahieren und erzielen auf neuen Daten schlechte Ergebnisse (vgl. [LKA16]). Aus diesem Grund werden beim Training von Entscheidungsbäumen keine weiteren Knoten auf einem Pfad erzeugt, wenn nach der Einteilung der Trainingsmenge an einem Knoten mehr als die Hälfte zu einer Klasse gehören. Im AdaBoost Algorithmus werden in der Trainingsphase Gewichte an die Trainingsdaten verteilt. Wortabbilder, die falsch klassifiziert werden, bekommen ein größeres Gewicht. Mithilfe der Gewichtsverteilung wird das Training des Entscheidungsbaumes neu durchgeführt. Der Entscheidungsbaum legt dann mehr Wert auf die richtige Zuordnung der falsch klassifizierten Wörter mit hohen Gewichten (vgl. [FS96]).

Für das Training wird, wie oben schon erwähnt, für jedes einzigartige Wort eine Wortklasse erstellt. Dabei wird auch hier das Zipf’sche Gesetz diskutiert. Kritisch be-


Abbildung 3.2.4: Das erste Bild zeigt das ursprüngliche Wortabbild. In der Mitte ist das künstlich erzeugte Wortabbild. Das letzte Bild zeigt den Unterschied der beiden Wortabbilder, indem diese überlappt werden. Bild entnommen aus [HRM05].

trachtet wird unter anderem, dass für bestimmte Datensätze viele Wörter nur einmal bzw. selten vorkommen. Dies bekräftigt die oben aufgestellte These, Cluster mit wenigen Einträgen nicht zu entfernen, da diese einen großen semantischen Wert haben können. Beim Training kann das zu Problemen führen. Klassen, für die es nur ein Wortabbild gibt, sind damit unterrepräsentiert. Um dies zu umgehen führen die Autoren für solche Klassen künstliche Daten ein. Dafür werden Varianten von Wortabbildern erzeugt, die sich leicht von dem ursprünglichen Wortabbild unterscheiden. Abbildung 3.2.4 zeigt das Wortabbild „Letter“ und eine künstlich erzeugte Version dessen.

### 3.2.5 *Segmentierungsfreies Word Spotting*

Der folgende Abschnitt befasst sich mit einigen Ansätzen für das segmentierungsfreie Word Spotting. Rusiñol et al. haben ein Patch-basiertes Vorgehen hierfür präsentiert [RATL15, RATL11]. In einem Patch-basierten Verfahren wird das zu analysierende Bild in regelmäßige Bildausschnitte unterteilt. Ein Patch beschreibt also einen Bildausschnitt. Jeder Patch wird als möglicher Wortkandidat betrachtet. Die Wahl der Größe und der Dichte, an denen Patches ermittelt werden, wird abhängig von der Zeilenhöhe getroffen. Zudem müssen Patches dicht genug berechnet werden, um alle Wortpositionen der Seite abzudecken.

Für jeden dieser Wortkandidaten muss eine geeignete Merkmalsrepräsentation bestimmt werden. Dafür wird die Bag of Features Repräsentation verwendet (siehe Abschnitt 2.3). Als Bilddeskriptoren für die BoF Repräsentation werden SIFT Deskriptoren verwendet, die in einem dichten Gitter berechnet werden. Zu den Patches wird auf Basis der Bag of Features eine Spatial Pyramid berechnet. Eine ausführliche Erläuterung zur Generierung einer Spatial Pyramid Repräsentation ist in Kapitel 2 zu

finden. Damit in der Praxis eine Anfrage effizient bearbeitet werden kann, werden die Merkmalsvektoren der Patches mit einer Dimensionsreduktion anhand von Latent Semantic Indexing (LSI) in einen Unterraum projiziert, um die hochdimensionalen Vektoren zu verkleinern. Der Unterraum wird mithilfe einer Singulärwertzerlegung berechnet und die Singulärvektoren mit dem größten Singulärwert werden für die Transformation in den Unterraum gewählt (vgl. [DDF+90]). Zur Anfragezeit wird die Spatial Pyramid Repräsentation der Query berechnet und in den Unterraum projiziert. Anhand der Kosinus Distanz werden die besten Patches identifiziert und zurückgegeben.

Rothacker et al. präsentieren in [RRF13] ein Patch-basiertes Word Spotting System. Anstatt die Größe der Patches abhängig von der Struktur des Dokumentes einheitlich festzulegen, werden diese über die Abmessungen der Anfragebox bestimmt. Dies hat den Vorteil, dass für neue Datensätze keine erneute Analyse über die Struktur der Dokumente (Zeilenhöhe) nötig ist. Wenn die Größenverhältnisse gleicher Wörter in der betrachteten Dokumentensammlung nicht zu stark abweichen, können so alle relevanten Bildausschnitte gefunden werden. In [RRLF14] wird dieser Ansatz von den Autoren weiter ausgeführt. Um die große Anzahl an potentiellen Wortkandidaten in einem Patch-basierten Verfahren und der daraus resultierenden Komplexität fürs Word Spotting zu vermeiden, wird das Bild nicht in regelmäßige Bildausschnitte unterteilt. Stattdessen werden zu einer Anfrage grobe Bildbereiche gesucht, die der Anfrage möglicherweise ähnlich sind. Innerhalb dieser werden regelmäßige Bildausschnitte betrachtet.

Patch-basierte Systeme sind eine beliebte und oft genutzte Variante um segmentierungsfreies Word Spotting durchzuführen. Für weitere Arbeiten in diesem Bereich sei auf [RFB+13, RF15, AGFV12] verwiesen.

Rodríguez und Peronnin haben einen anderen Schritt zur Wortkandidatengenerierung gewählt [RSP09]. Anstelle eines Patch-basierten Ansatz wird ein Verfahren auf Basis der Generierung von Worthypothesen gewählt. Zu Beginn werden Zusammenhangskomponenten berechnet. Angenommen es gibt einen Distanzwert $d_t$, sodass alle Komponenten zusammengruppiert werden die einen kleineren Distanzwert haben, eine optimale Segmentierung erzeugen. Dann entsprechen Distanzwerte, die größer als $d_t$ sind, Leerzeichen zwischen Wörtern. Ein solcher Wert ist stark problemabhängig und muss sehr konservativ gewählt werden, um keine einzelnen Wörter zu verschmelzen. Aus diesem Grund schlagen die Autoren eine Übersegmentierung vor, die durch Bilden von möglichen Worthypothesen entsteht. Anhand mehrerer Distanzwerte wird die Segmentierung durchgeführt und angenommen, dass jeder der Hypothesen ein Wortkandidat sein könnte. Die Übersegmentierung sorgt dafür, dass möglichst viele Wortkandidaten gefunden werden, im Gegenzug bedeutet das aber

eine höhere Komplexität in der Weiterverarbeitung durch die hohe Anzahl an Wortabbildern.

Über die Höhe und Breite der Anfrage kann die Menge der potentiellen Wortvergleiche gesteuert werden, indem Hypothesen gar nicht erst betrachtet werden, die sehr wahrscheinlich nicht zur Anfrage passen. Sind Hypothesen viel kleiner bzw. größer als die Anfragebox, können diese für die Ähnlichkeitsbewertung verworfen werden. Ein weiterer linearer Klassifikator verwirft Hypothesen, welche mit hoher Wahrscheinlichkeit keinen holistischen Wörtern entsprechen.

## 3.3 SCHLUSSFOLGERUNG

Verschiedene Verfahren zur Textdetektion wurden in diesem Kapitel vorgestellt. Im Bereich der natürlichen Szenenbilder hat sich der MSER Detektor bewährt (siehe Unterabschnitt 3.2.1). Dies motiviert dazu, den MSER Detektor im Kontext von historischen Dokumenten zu benutzen.

Die Arbeit von [RM06] behandelt die Indizierung im segmentierungsbasierten Fall (siehe Unterabschnitt 3.2.2) und hat somit einen starken Bezug zu dieser Arbeit. Es wurde unter anderem eine sequentielle Darstellung der Wortabbilder auf Basis der in Unterabschnitt 3.2.2 beschriebenen Profile verwendet. Diese Darstellungen benötigen eine gute Normalisierung der Wortabbilder [RRF13]. Im Gegensatz dazu ist für die BoF Darstellung keine explizite Vorverarbeitung der Wortabbilder nötig. Um die sequentiellen Darstellungen miteinander vergleichen zu können, wurde dort der DTW Algorithmus genutzt. Ein Nachteil des DTW ist die Komplexität. Eine bewährte Darstellung zur Angabe von Laufzeitkomplexitäten ist die O-Notation (vgl. [PD08] Kapitel 5.5), welche im Folgenden zum Vergleich von Komplexitäten benutzt wird. Bei Sequenzen der Länge $m$ und $n$ ergibt sich für das DTW eine obere Schranke von $O(mn)$ [PKG11]. Das beste Ergebnis erreichten die Autoren mit dem agglomerativen Clusteringverfahren Ward's Linkage. Der Nachteil bei agglomerativen Algorithmen ist auch hier die Komplexität. Bei $n$ Vektoren ergibt sich eine obere Laufzeitschranke von $O(n^2)$ (vgl. [AR13] Seite 107). Im Gegensatz dazu besitzt der Lloyd Algorithmus auf Basis der Kosinus-Distanz (Unterabschnitt 2.4.2), welcher hier benutzt wird, eine Laufzeit von $O(nki)$, wobei $k$ der Anzahl der Cluster und $i$ der Iterationen entspricht. Da in der Regel $k$ und $i$ kleiner als $n$ sind, ist der Lloyd Algorithmus vergleichsweise weniger komplex.

In Unterabschnitt 3.2.3 und 3.2.4 wurden Verfahren zur Worterkennung vorgestellt, mit dem Ziel eine automatische Transkription durchzuführen. Das Problem von holistischen Worterkennern ist, dass diese eine Entscheidung über die zu vergebende

Transkription eines Wortabbildes treffen müssen. Diese Aufgabe ist schwieriger als die visuelle Ähnlichkeitsbewertung zwischen Wortabbildern. Bei den Vorgehen handeln es sich um überwachte Lernverfahren, welche eine große Anzahl an annotierten Trainingsdaten benötigen. Im Gegensatz dazu findet die Indizierung semi-überwacht statt. Die grundlegende Idee dabei ist es, den hohen Aufwand bei der Annotation von Trainingsdaten zu umgehen, indem nur ein kleinerer Teil der Daten annotiert wird. Der Klassifikator wird dann mit Trainingsdaten gelernt, für die nur teilweise eine Klassenzuteilung bekannt ist (vgl. [Zhu05]). Bei der Indizierung werden dafür zu Beginn die Wortabbilder in einem unüberwachten Clustering Schritt gruppiert. Danach werden die Cluster in einem überwachten Schritt manuell annotiert [RVGF14, ASSM10]. Der Annotationsaufwand bei der Indizierung spiegelt sich somit in der manuellen Annotation der Cluster wider. Im Vergleich zu den beiden überwachten Verfahren ist dieser wesentlich geringer.

Am Ende des Kapitels wurden Methoden für das segmentierungsfreie Word Spotting vorgestellt. In [RATL15, RATL11] wurde die Arbeit präsentiert, nach der sich die Merkmalsdarstellung in dieser Arbeit richtet. Ein etabliertes Vorgehen im segmentierungsfreien Fall sind Patch-basierte Systeme [RRF13, RF15, RP08, RATL15, RATL11]. Dabei sind aber in der Regel Informationen über die Struktur des Dokumentes nötig, um die Patchgrößen zu bestimmen. Im klassischem Query-by-Example Word Spotting kann die Patchgröße über die Größe der Anfragebox ermittelt werden [RRLF14]. In der segmentierungsfreien Indizierung werden keine expliziten Anfragen gestellt. Dies hat zur Folge, dass die Größe der Patches nicht durch eine Anfragebox angenähert werden kann. Ein Patch-basiertes Verfahren müsste hier verschiedene Skalierungen verwenden, damit alle Wörter des Textes auch in einem Patch abgedeckt werden. Die Komplexität des Gesamtverfahrens steigt somit durch die große Anzahl an zu betrachtenden Bildausschnitten. Anstelle eines Patch-basierten Ansatz wird für die Textdetektion hier ein hypothesenbasiertes System angewendet, vergleichbar mit [RSP09]. Dabei werden statt Zusammenhangskomponenten, MSER kombiniert um Worthypothesen zu erzeugen. Der Vorteil gegenüber Zusammenhangskomponenten ist, dass keine feste Wahl auf einen Schwellwert zur Binarisierung nötig ist. Stattdessen werden bei dem MSER Verfahren alle möglichen Schwellwerte zur Binarisierung verwendet. Für die Textdetektion ist es also nicht nötig, einen optimalen Schwellwert zur Binarisierung zu wählen. Im Unterschied zu Patch-basierten Vorgehen sollen so Informationen über mögliche Textbereiche in die Berechnung von Wortkandidaten einfließen.

# 4

# METHODIK

Das folgende Kapitel befasst sich mit der vorgeschlagenen Methodik dieser Arbeit. Wie in den vorherigen Kapiteln festgestellt, ist die segmentierungsfreie Indizierung ein neuer Beitrag für die Dokumentenanalyse im Bereich der historischen Dokumente (vgl. Kapitel 1). Im Folgenden soll das Verfahren hierfür präsentiert werden. Es sei hier angemerkt, dass der prinzipielle Verlauf der Methodik für den segmentierungsbasierten und -freien Fall gleich bleibt. Daher wird das Kapitel sich auf die segmentierungsfreie Indizierung beziehen und an Stellen, an denen Unterschiede existieren, diese explizit erwähnen. Wenn im Folgenden Worthypothesen erwähnt werden, ist im segmentierungsbasierten Fall analog von Wortabbildern die Rede. Abbildung 4.0.1 zeigt den generellen Ablauf des Systems.

Falls der Kontrast innerhalb eines Dokumentes schwach ist, wird eine Vorverarbeitung des Dokumentes durchgeführt (Abschnitt 4.1). Hiernach wird die Textdetektion mit einem hypothesenbasierten Ansatz beschrieben (Abschnitt 4.2), um potentielle Wortregionen zu finden. Im segmentierungsbasierten Fall ist die Segmentierung durch die Annotation gegeben. Für jeden dieser Wortkandidaten wird ein Merkmalsvektor in Form einer Spatial Pyramid berechnet. Dies wird in Abschnitt 4.3 beschrieben. Nachdem für jeden Wortkandidaten eine Merkmalsdarstellung bestimmt wurde, wird in Abschnitt 4.4 das Clustering für die Gruppierung von visuell ähnlichen Wörtern ausgeführt. Zum Schluss werden die entstandenen Cluster annotiert.

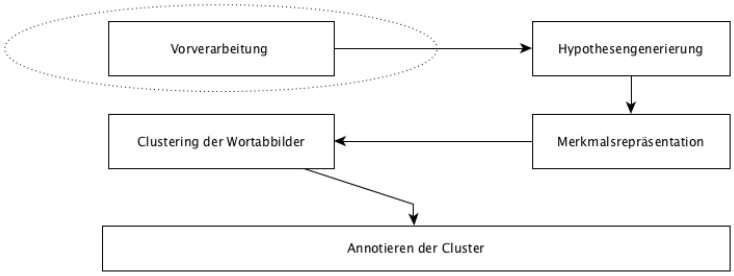



Abbildung 4.0.1: In der Abbildung ist der Ablauf der vorgestellten Methodik dargestellt. Zu Beginn wird, falls nötig eine Vorverarbeitung der Dokumente durchgeführt. Danach gilt es die Wörter der Dokumente zu finden. Hierauf aufbauend wird für jede Hypothese eine Merkmalsdarstellung berechnet. Zum Schluss werden diese geclustert, um visuell ähnliche Wörter zu gruppieren. Zuletzt werden die Cluster annotiert.

## 4.1 VORVERARBEITUNG

Dokumente werden im Computer als Graustufenbilder interpretiert. Dabei kann es vorkommen, dass durch das Altern der Dokumente, der Kontrast zwischen Schrift und Hintergrund schwach ist. Abbildung 4.1.1 zeigt ein kontrastreiches Dokument. Im Gegensatz dazu sind dort Briefe dargestellt, die wesentlich kontrastärmer sind. Aus diesem Grund werden kontrastarme Bilder in einem Vorverarbeitungsschritt angepasst. Hierfür wird die Methodik verwendet, die auch in [FRG14] von Fink et al. vorgeschlagen wurde. Ein Graustufenbild ist auch als Histogramm über die Häufigkeiten der Grauwerte darstellbar. Kontrastarme Bilder zeichnen sich dadurch aus, dass es einen Punkt im Histogramm gibt, an denen sich der Großteil der Pixelwerte sammeln und zudem nicht der gesamte Wertebereich verwendet wird. Mit einem Histogrammausgleich wird das Histogramm flacher und die Intensitäten verteilen sich gleichmäßiger auf den Wertebereich. Dies führt zu einer Kontrasterhöhung (vgl. [GW06] Kapitel 3.3). Um Hintergrundrauschen und -Artefakte zu entfernen wird das Bild zusätzlich mithilfe eines Medianfilters bearbeitet. Der Medianfilter ist ein Rangordnungsfilter. Für jeden Pixel werden zusätzlich die Pixel betrachtet, die innerhalb der definierten rechteckigen Maske liegen. Die relevanten Pixel werden aufsteigend sortiert und der Median gebildet, welcher als neuer Pixelwert festgelegt wird (vgl. [Jä12] Kapitel 11.6.1). Abbildung 4.1.2 (b) zeigt das Dokument nach der Histogrammglättung und dem Anwenden des Medianfilters.

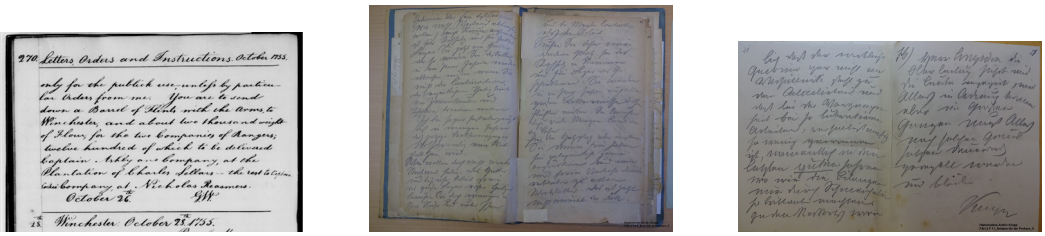
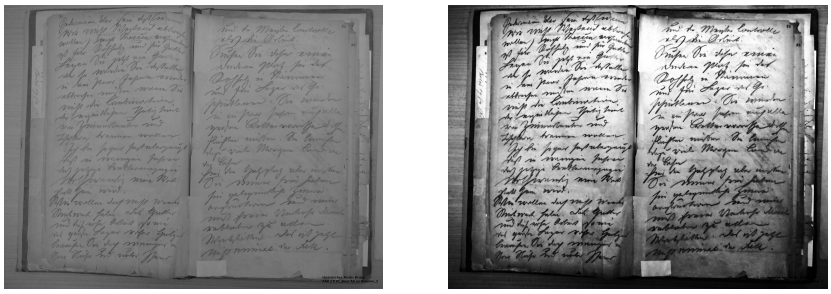


Abbildung 4.1.1: Links befindet sich ein Ausschnitt eines kontrastreiches Dokumentes. Die anderen beiden Bilder sind Briefe, die einen eher schwachen Kontrast aufweisen.



(a) (b)

Abbildung 4.1.2: a) Unverarbeiteter Dokumentenbrief. Auf dem Bild ist der Kontrast zwischen Schrift und Hintergrund kaum vorhanden. b) Bild nach Anwendung der Histogrammglättung und des Medianfilters. Der Kontrast wurde erhöht.

## 4.2 TEXTDETEKTION UND HYPOTHESENGENERIERUNG

Unabhängig davon ob eine Vorverarbeitung stattgefunden hat, gilt es im Weiteren die Wörter im Bild zu identifizieren. Im segmentierungsbasierten Fall wird die Segmentierung aus der Annotation übernommen. Die Idee der hypothesenbasierten Textdetektion ist es kleinere Textstellen, wie einzelne Buchstaben und Teilwörter zu größeren Strukturen zusammenzufügen, die möglichst mit den Wörtern der Dokumente korrespondieren.

### 4.2.1 *MSER Detektion zum Finden von Textregionen*

Für die Generierung von Worthypothesen müssen zunächst kleinere Textregionen identifiziert werden, die dann zu Hypothesen kombiniert werden. Zum Finden möglicher Wortstrukturen wird der in Kapitel 2 beschriebene MSER Detektor genutzt. In der im Unterabschnitt 3.2.1 beschriebenen Arbeit [NM11a] wurde das MSER Verfahren auch auf den einzelnen Rot, Grün und Blau Kanälen durchgeführt. Dies ist sicherlich sinnvoll im Kontext der natürlichen Szenen, in denen Text in unterschiedlichen Farben auftauchen kann. Im Bereich der historischen Dokumente ist aber davon auszugehen, dass Schriftzüge eine einheitliche Farbe (schwarze Töne) besitzen und auf Papier geschrieben wurden (weiße Töne). Somit werden die meisten historischen Dokumente mit Grauwertintensitäten kodiert. Deswegen wird an dieser Stelle der MSER Detektor auf Graustufenbilder verwendet. Die Implementierung des MSER Detektors wird der OpenCV Bibliothek [Braoo] entnommen. Folgende Parameter sind in dieser einstellbar:

- **Delta $\Delta$:**

  Der Parameter $\Delta$ beeinflusst die Berechnung der Stabilität einer Zusammenhangskomponente $z$. Bei dem MSER Verfahren wird das Bild anhand allen möglichen Schwellwerten binarisiert. In Abbildung 4.2.1 ist exemplarisch ein MSER-Baum für einige Binarisierungsschwellwerte dargestellt. $\Delta$ bestimmt für wie viele Schwellwerte eine Region stabil sein muss. Zu einem Schwellwert $i$ werden zusätzlich die $i - \Delta$ Schwellwerte betrachtet. Je größer also $\Delta$ gewählt wird, desto mehr Schwellwerte gehen in die Berechnung der Flächenänderung einer Region ein. Die Fläche $A$ einer Komponente ist als Anzahl ihrer Pixel definiert. Dann berechnet sich die Variation $V$ einer Region mit :

  $$V = \frac{A(z)_i - A(z)_{i-\Delta}}{A(z)_{i-\Delta}} \tag{4.2.1}$$

  Je kleiner die Variation einer Region, desto stabiler ist sie.

- **Maximale Variation:**

  Dieser Parameter gibt an, wie stark sich eine Region über die durch $\Delta$ festgelegten Schwellwerte ändern darf. Der anzugebende Wert liegt zwischen $0 - 1$ und je größer er ist, desto mehr Regionen werden berechnet. Ist die Flächenänderung einer Region über die $\Delta$ Schwellwerte zu groß, kann diese verworfen werden. In Abbildung 4.2.1 ist dies mit dem grünen Pfeil gekennzeichnet (2). Der Flächenunterschied der Komponente zur Elternkomponente ist zu groß und wird verworfen.

- **Minimaler Unterschied:**

  Innerhalb eines Bildausschnittes werden über Folgen von Schwellwerten mehrere Regionen gefunden, die sich nur um wenige Pixel unterscheiden. Damit nicht zu viele wiederholende Regionen berechnet werden, entscheidet der minimale Unterschied darüber, wie sehr diese Pixelanzahl abweichen muss, damit diese nicht verworfen werden. In Abbildung 4.2.1 ist dies an dem blauen Pfeil dargestellt (3).

- **Minimale und Maximale Fläche der MSER:**

  Bei der verwendeten Implementierung ist es möglich MSER, welche zu groß oder zu klein sind, zu entfernen. Betrachtet wird hierfür die Fläche $A$ einer MSER. Ist diese kleiner bzw. größer als die gewählten Schranken, werden diese entfernt.

Jede einzelne MSER wird als Menge seiner enthaltenden Pixel zurückgegeben. Für die Weiterverarbeitung wird jede MSER durch das kleinste umrahmende Rechteck repräsentiert, welches alle Pixel enthält. Abbildung 4.2.2 visualisiert beispielhaft berechnete MSER auf einem Ausschnitt eines Dokumentes.

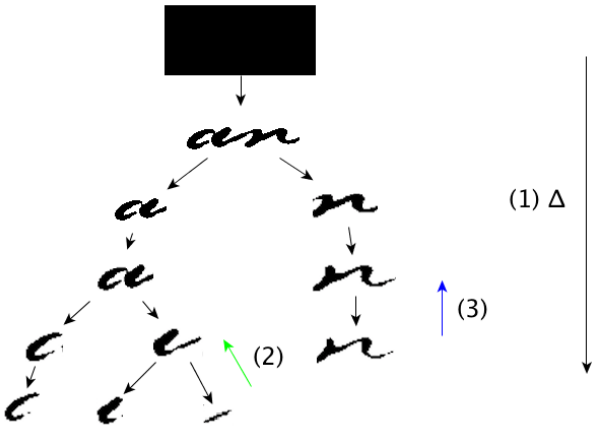



Abbildung 4.2.1: In der Abbildung ist ein MSER-Baum für ausgewählte Schwellwerte zur Binarisierung dargestellt. Auf jeder Ebene des Baumes sind die Zusammenhangskomponenten des aktuellen Schwellwertes visualisiert. Verschmelzen Komponenten bei einem höheren Schwellwert, werden diese unter einem Elternknoten zusammengefasst. (1) zeigt den Parameter $\Delta$. Mit dem grünen Pfeil (2) ist ein Fall dargestellt, in dem die maximale Variation einer Komponente zu hoch ist. (3) stellt den Parameter für den minimalen Unterschied dar.

### 4.2.2 *Filtern der Regionen*

Die Menge der MSER beinhalten unter anderem Bereiche, die keine Wortteile enthalten (siehe Abbildung 4.2.2). Aus diesem Grund werden hier MSER, die mit großer Wahrscheinlichkeit keinen Textbereichen entsprechen, entfernt. Dafür wird anhand einiger geometrischer Eigenschaften die Menge der Regionen gefiltert. Besonders lange oder breite Boxen werden verworfen, da diese eher Artefakte beinhalten. Zu beobachten sind diese vor allem an den Rändern des Dokumentes (siehe Abbildung 4.2.2).

### 4.2.3 *Hypothesengenerierung*

Die detektierten MSER Bereiche entsprechen im Optimalfall Textbereichen. Diese gilt es zu Worthypothesen zu kombinieren. Der Algorithmus zur Hypothesengenerierung geht dabei wie folgt vor: Sei $M$ die Menge der als Textregionen identifizierten MSER. Zu jeder Region $m \in M$ werden alle möglichen Kombinationen mit Elementen aus $M$ gebildet, die innerhalb der Nachbarschaft von $m$ liegen. Die Nachbarschaft von $m$ lässt sich auffassen als die MSER, die nicht zu weit von $m$ entfernt sind. Als Entfernung zweier Boxen wird hier die minimale Distanz in Pixeln sowohl in x- als auch y-Richtung betrachtet. Damit Wortregionen, welche zu weit auseinander liegen, nicht kombiniert werden, werden maximale Distanzwerte festgelegt. Im Weiteren werden diese als $d_x$ und $d_y$ bezeichnet. Mit den hierdurch erzeugten Hypothesen werden weitere Hypothesen generiert. Dafür werden nach dem gleichen Prinzip die neuen Hypothesen, unter Berücksichtigung der maximalen Distanzwerte, mit den ursprünglichen MSER kombiniert. Dies wird wiederholt, bis keine weiteren Hypothesen mehr erzeugt werden.

Die Wahl der Distanzwerte $d_x$ und $d_y$ ist hierbei entscheidend. Werden diese zu konservativ gewählt, werden zu wenig Hypothesen generiert und somit weniger Wörter erfasst. Bei zu großen Werten für die Distanzwerte werden zusätzliche Hypothesen erzeugt, ohne dabei die Detektionsrate zu verbessern. Abbildung 4.2.3 verdeutlicht das wesentliche Konzept bei der Hypothesengenerierung. Wenn Regionen zu weit auseinander liegen, werden mit diesen keine neuen Wortvarianten erzeugt.

## 4.3 REPRÄSENTATION DER WORTABBILDER

Nachdem die Worthypothesen generiert worden sind, gilt es für diese Bildbereiche Merkmalsrepräsentationen zu bestimmen. Fundamental sind hierfür die Konzepte des SIFT Deskriptors, Bag of Features und Spatial Pyramid und es sei an dieser Stelle

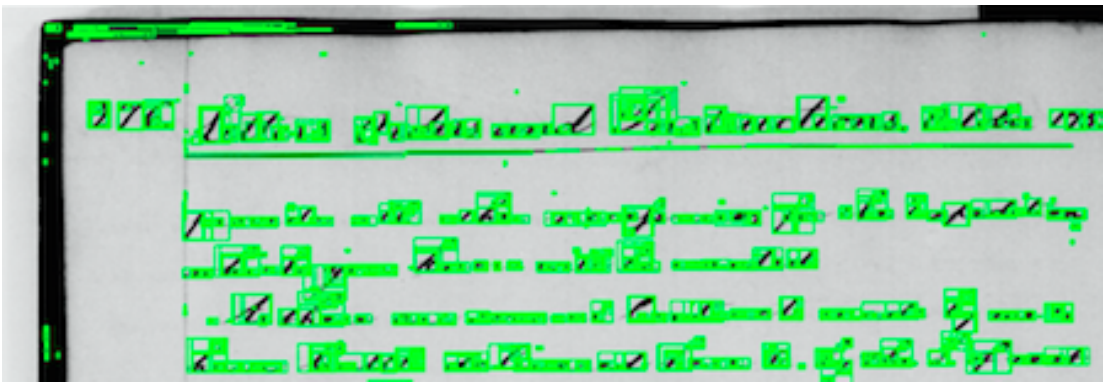


Abbildung 4.2.2: MSER Regionen auf einem Ausschnitt des Dokumentes. Kleinere Wortstrukturen werden gefunden. Zu beachten ist hierbei, dass zusätzlich MSER generiert werden, die nur Hintergrund beinhalten. Wie an den schwarzen Ränder zu erkennen werden auch hier MSER berechnet.

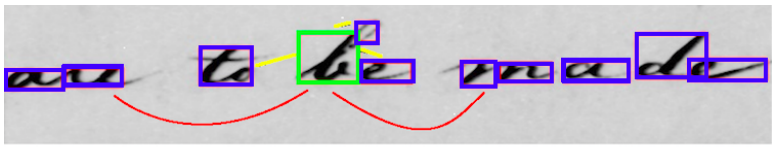


Abbildung 4.2.3: Beispiel zur Hypothesengenerierung: Zu einer MSER (grün) werden die Regionen gesucht, die innerhalb eines definierten Distanzwertes liegen (gelbe Pfeile). Mit diesen werden Worthypothesen erzeugt. Die roten Pfeile deuten auf Komponenten hin, die zu weit auseinander entfernt sind und zu unterschiedlichen Wörtern gehören und es werden mit diesen keine neuen Wortvarianten generiert.

auf Kapitel 2 verwiesen. Abbildung 4.3.2 visualisiert die Merkmalsberechnung für Worthypothesen. Zu Beginn werden auf jedem Dokument SIFT Deskriptoren in einem dichten Gitter berechnet (1.). Diese werden mithilfe des Lloyd Algorithmus auf Basis der euklidischen Distanz geclustert (2.). Die berechneten Zentroiden am Ende des Algorithmus repräsentieren charakteristische Deskriptoren und werden als Kodebuch für die Bag of Features Darstellung gewählt.

Worthypothesen sind durch ein Rechteck eingegrenzt. Für den Merkmalsvektor eines Wortes werden die Deskriptoren gewählt, die innerhalb des zugehörigen Rechteckes liegen. SIFT Deskriptoren, die zu weit am Rand des Rechteckes liegen, betrachten unter anderem Pixel, die außerhalb dieses Rechteckes liegen. Somit besteht die Möglichkeit, dass der Deskriptor Schriftverläufe von anderen naheliegenden Wör-

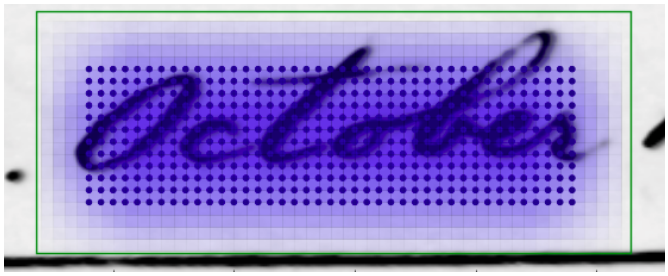


Abbildung 4.3.1: Auswählen der Deskriptoren eines Wortabbildes. Die schwarzen Punkte stellen die Mittelpunkte der Deskriptoren dar. Mit der blauen Färbung sind die Zellen der Deskriptoren visualisiert. Je dunkler eine Fläche ist, desto mehr Deskriptoren überdecken diese. Deskriptoren, die über den Rand des Bildausschnittes hinausgehen, werden verworfen.

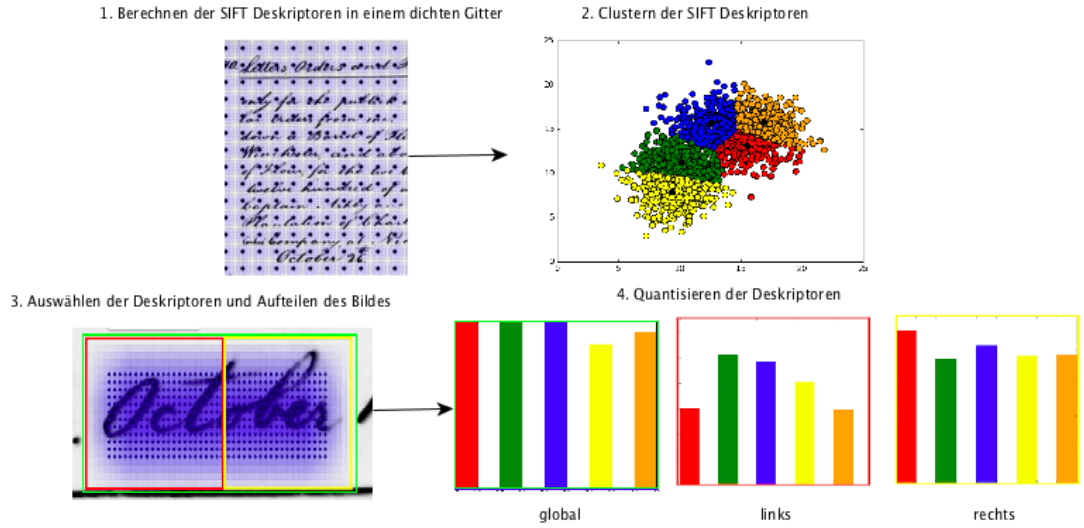



Abbildung 4.3.2: Die Berechnung der Spatial Pyramid Darstellung eines Wortes. Zu Beginn werden auf allen Dokumenten SIFT Deskriptoren in einem dichten Gitter berechnet (1.). Danach werden diese mit dem Lloyd Algorithmus zu einem Kodebuch geclustert (2.). Für jedes Wortabbild werden dann die zugehörigen Bilddeskriptoren mit dem Kodebuch quantisiert und in jedem Teilbild (grün,rot,gelb) eine BoF Repräsentation berechnet. Zum Schluss werden diese für die Spatial Pyramid Darstellung konkateniert.

tern mit einbezieht. Aus diesem Grund werden Deskriptoren verworfen, die über den Rand der zugehörigen Box hinausragen (siehe Abbildung 4.3.1).

Sind zu einem Wortabbild die entsprechenden SIFT Deskriptoren ausgewählt worden, kann die Spatial Pyramid Repräsentation eines Wortes berechnet werden. Die SIFT Deskriptoren werden anhand des generierten Kodebuches quantisiert. D.h. in diesem Fall wird zu jedem SIFT Deskriptor der nächste Nachbar aus dem Kodebuch bestimmt und ihm zugewiesen. Entsprechend der Spatial Pyramid Aufteilung wird das Bild unterteilt und innerhalb jedes Teilbildes ein Histogramm über die auftretenden Einträge des Kodebuches gebildet. Die berechneten Bag of Features Repräsentationen ergeben konkateniert die Spatial Pyramid (3. und 4.). Innerhalb des generierten Merkmalsraum können jetzt Wortabbilder im Clustering Algorithmus auf Ähnlichkeit bewertet werden.

## 4.4 CLUSTERING VON WORTABBILDERN

Im Folgenden wird das Clustering von Wortabbildern zur Generierung einer Transkription näher beschrieben. Nachdem für die Wortabbilder Merkmalsvektoren berechnet worden sind, ist der nächste Schritt ähnliche Wortabbilder in Clustern zu gruppieren. Da die berechneten Merkmalsdarstellungen sehr hochdimensional sind, wird für diesen Schritt der Lloyd Algorithmus auf Basis der Kosinus-Distanz angewendet (vgl. Unterabschnitt 2.4.2). Die euklidische Distanz ist hierfür nicht geeignet, da Unterschiede in einzelnen Dimensionen durch das Quadrieren zu stark ins Gewicht fallen. Die Merkmalsvektoren werden für die Berechnung der Kosinus-Distanz L2-normalisiert. Als Startkodebuch werden aus der Gesamtmenge der Worthypothesen zufällig $k$ Vektoren gewählt. Das $k$ sollte hierbei idealerweise so gewählt werden, dass für jedes verschiedene Wort genau ein Cluster entsteht. Dieser Wert ist in der Praxis nicht bekannt und wird geschätzt. Eine Möglichkeit hierfür ist das Gesetz von Heap (vgl. Unterabschnitt 3.2.2). In Kapitel 5 wird gezeigt, dass die zufällige Initialisierung der Zentroiden, mit einer ausreichenden Anzahl an Iterationen, nur einen kleinen Einfluss auf das Endergebnis hat.

Am Ende des Lloyd Algorithmus entsteht ein Kodebuch, in dem die Kodes den Repräsentanten der $k$ Clustern entsprechen. Ein Zentroid ist zwar der Repräsentant eines Clusters, gehört aber im Regelfall nicht zu der Menge der Wortabbilder selbst. Bevor zu jedem Cluster das Label bestimmt werden kann, muss das repräsentative Wortabbild eines Clusters gefunden werden. Dafür werden im nächsten Schritt die Medoiden, die nächsten Nachbarn im Merkmalsraum, zu jedem Zentroid des Kodebuches berechnet.

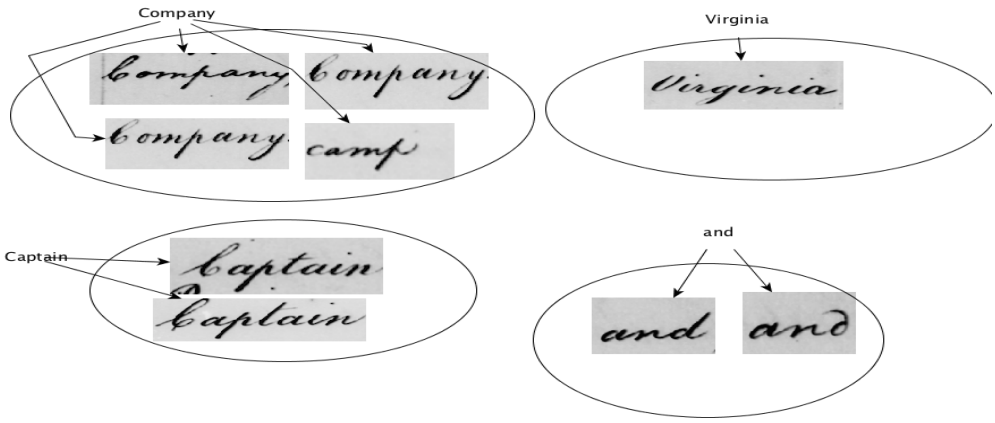



Abbildung 4.4.1: Nachdem die Annotation für die Cluster bestimmt wurden, werden diese an die Wortabbilder propagiert. In diesem Beispiel sind vier Cluster dargestellt. Die Cluster wurden mit „Company“, „Captain“, „and“, „Virginia“ annotiert. Das Propagieren der Clusterannotation an die Wortabbilder des Clusters führt zu einer kompletten Transkription der Dokumentensammlung.

Sei $c_i$ ein Zentroid und $S$ die Gesamtheit aller Spatial Pyramid Darstellungen der Worthypothesen. Dann werden die Medoiden anhand folgender Gleichung ermittelt:

$$m_i = \underset{\forall s_i \in S}{\arg\min}(1 - \cos(s_i, c_i)) \tag{4.4.1}$$

Zu jedem Clusterzentroid wird die Worthypothese des jeweiligen Clusters bestimmt, welches die kleinste Kosinusdistanz zu ihm besitzt. Der Medoid eines Clusters ist somit das Wortabbild, welches den Cluster am besten repräsentiert. Für jeden Cluster wird durch den berechneten Medoiden das Clusterlabel bestimmt. In der Praxis würde hier der in Abschnitt 3.3 beschriebene Annotationsaufwand anfallen. An dieser Stelle vergibt ein Nutzer anhand des Aussehens des Clustermedoiden eine Transkription. Dieser Schritt wird in dieser Arbeit simuliert, indem innerhalb der Annotation das Wortabbild gesucht wird, welches die größte Überlappung zu dem Medoiden hat. Im segmentierungsbasierten Fall genügt es hier, die Annotation für den Medoiden zu bestimmen.

Zu jedem Cluster ist jetzt ein Label zugewiesen worden. Abbildung 4.4.1 visualisiert den nächsten Schritt. Die Clusterlabel werden an die Wortabbilder, die innerhalb ihres Clusters liegen, weiterpropagiert. Die Cluster, durch die Ellipsen dargestellt, wurden mit „Company“, „Captain“, „and“, „Virginia“ annotiert. Jedes Wortabbild

wird mit dem zugehörigen Clusterlabel transkribiert. Dies kann wie im Fall des 1. Clusters zu Fehlerfällen führen, wenn ein Wortabbild dem falschen Cluster zugeordnet wird. Das Wort „camp“ liegt im Cluster mit der Annotation „Captain“ und wird somit falsch erkannt. Die Qualität der Erkennungsrate hängt somit von der Güte des Clustering ab.

# 5 EVALUIERUNG

Nachdem die verwendete Methodik im vorherigen Kapitel präsentiert wurde, sollen in diesem Kapitel die Experimente vorgestellt werden, die durchgeführt wurden, um die Leistungsfähigkeit des vorgeschlagenen Verfahrens zu testen. Für die Auswertung sind Datensätze historischer Dokumente nötig, die in Abschnitt 5.1 vorgestellt werden. In Abschnitt 5.2 werden die Metriken erläutert, die im Weiteren für die Evaluierung benutzt werden. Abschnitt 5.3 beschreibt das verwendete Evaluierungsprotokoll. Abschnitt 5.4 präsentiert die Ergebnisse der Evaluation. Zu Beginn wird die Textdetektion und anschließend die segmentierungsbasierte Indizierung ausgewertet. Mit den besten gefundenen Parameterkonfigurationen werden die einzelnen Teile zur segmentierungsfreien Indizierung zusammengeführt. Zum Schluss wird ein Vergleich zu den Ergebnissen des aktuellen Standes der Technik zur Worterkennung gegeben (Abschnitt 5.5).

## 5.1 DATENSÄTZE

In diesem Abschnitt wird auf die Datensätze eingegangen, die im Hinblick der Evaluierung gewählt wurden. Die Auswertung findet auf zwei Datensätzen statt, die aus historischen handgeschriebenen Dokumenten bestehen. Hierbei soll auf die jeweiligen Eigenschaften und Rahmenbedingungen eingegangen werden. Zu Beginn wird der George Washington Datensatz vorgestellt. Mit dem Krupp Datensatz wird ein neuartiger Datensatz präsentiert.

### 5.1.1 *George Washington Benchmark*

Der George Washington Datensatz ist einer der weit verbreiten Benchmarks in der Dokumentenanalyse [RRF13, RF15, RRLF14, RATL15, KWD14, RM06, LRM04, MR05]. Er umfasst 20 gescannte Seiten der George Washington Sammlung des Library of Congress[1]. Es handelt sich bei den Dokumenten dieses Datensatzes um handschriftlich verfasste englischsprachige Texte. Zu dem Datensatz existiert eine manuell angefertigte Annotation bestehend aus den Rechtecken, welche die Wortabbilder eingrenzen

1 Library of Congress: https://memory.loc.gov/ammem/gwhtml/

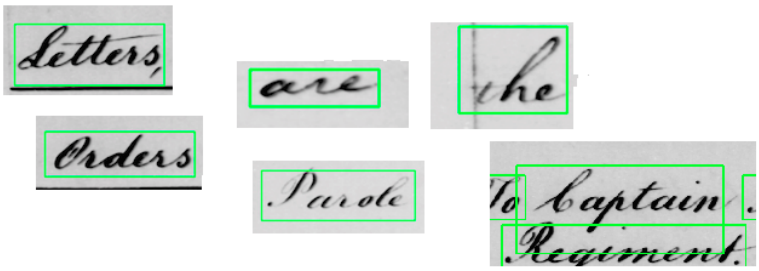



Abbildung 5.1.1: Exemplarisch einige mit der Annotation segmentierte Wortabbilder des GW Datensatz. Zu Problemen führen Wortabbilder, die Schriftpixel anderer Wörter beinhalten. Im Wortabbild „Captain“ sind Schriftverläufe von den umliegenden Wörtern „to“ und „Regiment“ enthalten.

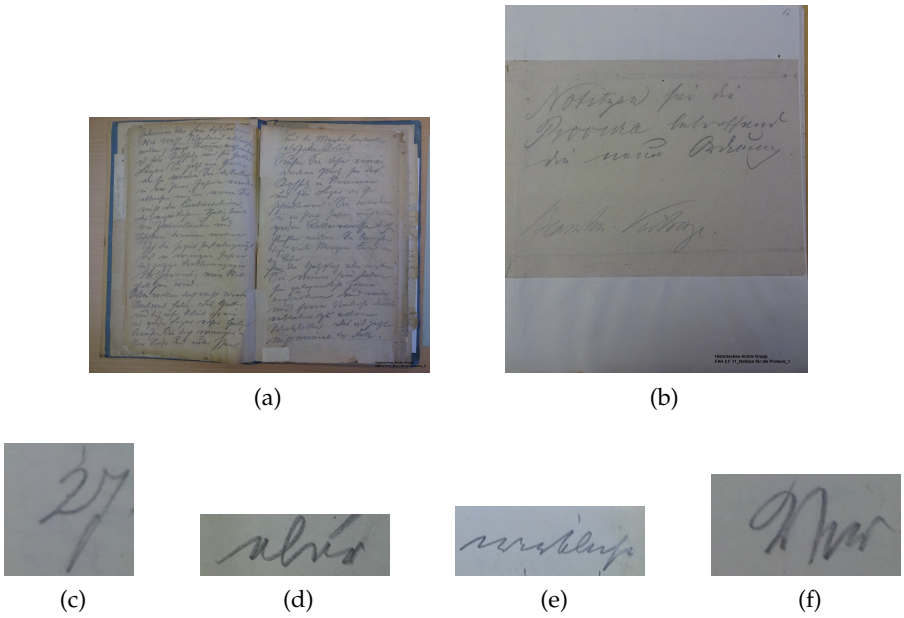
und deren zugehörigen Transkriptionen. Insgesamt besteht der Datensatz aus 4860 Wortabbildern mit 1187 verschiedenen Wörtern [RM06]. Mit verschiedenen Wörtern sind hierbei die unterschiedlichen Wortklassen gemeint. In Abbildung 5.1.1 sind beispielhaft einige Wortabbilder dargestellt.

Anzumerken ist hierbei, dass die Annotationen an einigen Stellen unpräzise sind und Schriftverläufe anderer Wörter beinhalten. Zudem sind die Transkriptionen von einigen Wörtern nicht einheitlich. Für das Datum 28. finden sich zum Beispiel die Annotation „28th“ und „28“. Im weiteren Verlauf wird der Datensatz mit GW abgekürzt.

### 5.1.2 *Krupp Briefe*

Für die Arbeit wurde ein Teil der Briefe aus dem Krupp Archiv[2] manuell transkribiert. Bei den Dokumenten handelt sich um Geschäftsbriefe des Unternehmers Alfred Krupp an einige Geschäftspartner. Insgesamt wurden 12 Seiten der Sammlung annotiert. Der angefertigte Datensatz beinhaltet 766 Wortabbilder mit 446 verschiedenen Wörtern. Die Dokumente sind in fotografierter statt gescannter Form zur Verfügung gestellt worden. Dies hat zur Folge, dass auf einigen Seiten im Hintergrund zusätzliche Objekte, die nicht zum Dokument gehören, dargestellt sind. In Abbildung 5.1.2 (a) ist ein Dokument des Krupp Datensatzes abgebildet, in dem im Hintergrund, Oberflächen eines Tisches zu sehen sind. Zusätzlich ist in Abbildung 5.1.2 (b) ein weiteres Problem zu sehen. Die Größenunterschiede von Wörtern variieren auf einigen Seiten stark.

2 Historisches Archiv Krupp, Essen, Deutschland



(a) (b)

(c) (d) (e) (f)

Abbildung 5.1.2: (a) Dokument des Krupp Datensatzes. Im Hintergrund sind Umrisse eines Tisches zu sehen. Die Qualität des Dokumentes ist durch das Verbleichen des Papiers stark beeinträchtigt. (b) In der Sammlung herrschen zum Teil größere Unterschiede in der Größe der Wortvorkommen. (c)-(f) Einige segmentierte Wortabbilder des Krupp Datensatzes.

Abbildung 5.1.2 (c)-(f) zeigen einige Wortabbilder des Datensatzes. Eine weitere unmittelbare Schwierigkeit ist, dass die Briefe in der alten deutschen Schriftart Kurrent handschriftlich verfasst worden sind. Kurrent zeichnet sich in seiner Form durch sehr ähnliche Schriftverläufe aus. Für unerfahrene Leser sind die unterschiedlichen Buchstaben kaum auseinander zu halten. Verstärkt wird dieser Effekt durch die kontrastarme Schrift und dem Verbleichen der Dokumente (siehe Abbildung 5.1.2 (a)). Bei den Fotos der Dokumente handelt es sich um Farbbilder, die für den weiteren Verlauf in Graustufenbilder transformiert werden. Im Folgenden wird der Datensatz als Krupp Datensatz bezeichnet.

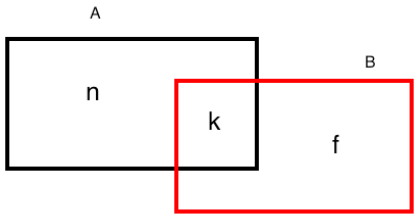
## 5.2 METRIKEN

Der folgende Abschnitt befasst sich mit den verwendeten Maßen, die sowohl zur Bewertung der Güte der einzelnen Komponenten der Methodik als auch zum Vergleich mit dem Stand der Technik dienen.

### 5.2.1 *Intersection over Union*

Eine Metrik zur Qualitätsbewertung einer einzelnen Detektion ist die Intersection over Union (IoU) [SNL14]. Sie gibt den Überdeckungsgrad einer Detektion A zu einem Element B aus der Annotation an. Betrachtet werden dabei die durch A und B definierten Bildflächen. Dann lässt sich dieser Wert über die Gleichung 5.2.1 berechnen.

$$\mathrm{IoU} = \frac{|A \cap B|}{|A \cup B|} = \frac{|k|}{|n| + |f| - |k|} \tag{5.2.1}$$




Der Wert entspricht somit dem Verhältnis zwischen der Schnittmenge (korrekt erkannte Fläche) und der Vereinigung beider Flächen (korrekt erkannte + falsch erkannte + nicht erkannte Fläche). Im Fall von Rastergrafiken reicht es aus, die Pixel innerhalb der Schnittmenge und der Vereinigung zu bestimmen. Die Werte für das Maß liegen im Bereich von 0 bis 1, wobei größere Werte eine bessere Überdeckung der Annotation bedeuten. In der Regel gilt ein Objekt als erkannt, wenn es eine Detektion gibt, die einen IoU Wert von mindestens 0.5 zu diesem hat.

### 5.2.2 *Detektionsrate*

Für die Textdetektion ist es wichtig den prozentualen Anteil an lokalisierten Wörtern in der Dokumentensammlung zu bestimmen. Hierfür wird die Detektionsrate (DR) eingeführt. Mit Gleichung 5.2.2 wird die DR berechnet. Dabei bezeichnet N die Anzahl aller Wörter der Annotation und ein Wortabbild gilt als korrekt lokalisiert, wenn

eine Worthypothese existiert, die mindestens einen IoU Wert von 0.5 mit diesem bildet.

$$DR = \frac{|\text{korrekt lokalisierte Wörter}|}{N} \tag{5.2.2}$$

Diesen Wert gilt es zu maximieren, um möglichst viele Wörter der Dokumentensammlung zu finden und wird im Weiteren in Prozent angegeben.

### 5.2.3 *Worterkennungsrate*

Um die Qualität der Indizierung zu prüfen ist ein Maß nötig, welches die generierte Transkription zu einer gegebenen Annotation beurteilt. Ein einfaches Maß hierfür ist die Worterkennungsrate. Die Worterkennungsrate (WR) beschreibt den prozentualen Anteil an korrekt transkribierten Wörtern. Zu jedem Wortabbild gibt es eine manuell angefertigte Annotation, diese wird mit der automatisch erzeugten Transkription verglichen. Sind diese gleich, so gilt dieses Wortabbild als korrekt transkribiert. Daraus lässt sich dann die WR mit der Gleichung 5.2.3 ableiten. Auch hier ist N wieder die Anzahl aller Elemente der manuellen Transkription.

$$WR = \frac{|\text{korrekt transkribierte Wörter}|}{N} \tag{5.2.3}$$

Da die Ergebnisse der Indizierung im Bezug auf die WR stark von der Anzahl der Cluster und somit dem benötigten Annotationsaufwand abhängig sind, wird für diese Arbeit ein weiteres Maß definiert, die durchschnittliche Worterkennungrate. Diese wird im weiteren mit ∅WR abgekürzt. Ziel dieses Maßes ist es eine allgemeine Bewertung der Indizierung über verschiedene Annotationsaufwände zu geben. Gedacht ist das Maß vor allem für die Konfiguration der Parameter, um das Verfahren für unterschiedlich großen Annotationsaufwand zu optimieren. Er berechnet sich aus dem arithmetischen Mittel der WR über die verschiedenen Clusteranzahlen k. Dabei werden die k so gewählt, dass diese einem Vielfachen des Zehntels der Gesamtanzahl an Wörter des Datensatzes entspricht. Damit wird, wie in Abbildung 5.2.1 zu sehen, die Fläche unter der Kurve über alle möglichen k von Clustern angenähert.

### 5.2.4 $\text{Maß}_1$ *und* $\text{Maß}_2$

Für den segmentierungsfreien Fall ist die WR nicht ohne weiteres übertragbar, da Worthypothesen betrachtet werden. Würde die WR in Bezug zur Anzahl der Worthypothesen berechnet werden, gibt diese keine ausreichende Information über den

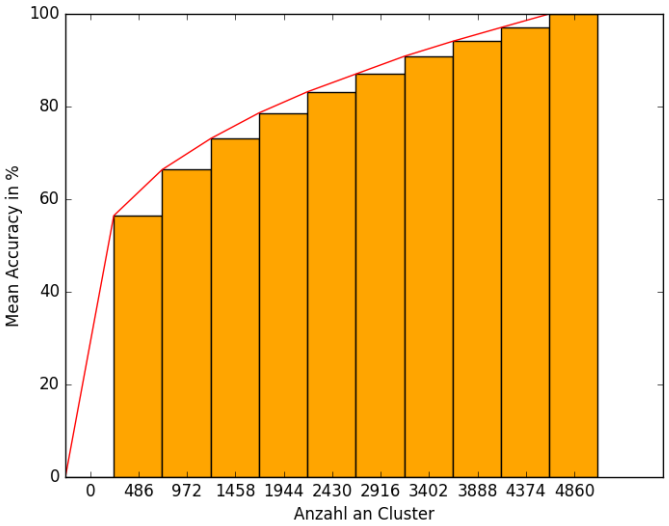



Abbildung 5.2.1: Die ∅WR bestimmt das arithmetische Mittel über die WR der verschiedenen Clusteranzahlen. Wie in der Abbildung zu sehen, wird so die Fläche unterhalb der WER Kurve für alle möglichen Werte von Kodebuchgrößen approximiert. Bei $k = 0$ und $k = n$ ist die WR trivialerweise immer 0% bzw. 100%.

Anteil an korrekt erkannten Wörtern der Annotation. Stattdessen wird überprüft, wie viele Wörter aus der Annotation richtig erkannt werden. Dafür wird getestet, ob es eine Hypothese gibt, die mit einem Wortabbild zu einem IoU Wert von mindestens 0.5 überlappt und richtig transkribiert wurde. Ein strengeres Maß wird mit $Maß_2$ definiert. Jedes Wortabbild aus der Annotation wird nur dann als korrekt erkannt gezählt, wenn es zusätzlich keine Hypothese gibt, die mit dieser überlappt und eine falsche Transkription zugewiesen bekommen hat.

## 5.3 PROTOKOLL

Im Folgenden wird das verwendete Evaluierungsprotokoll kurz vorgestellt. Alle Experimente werden grundsätzlich mit allen Seiten der Datensätze durchgeführt. Für den GW Datensatz sind das 20 Seiten (4860 Wortabbilder) und bei dem Krupp Datensatz 12 Seiten (766 Wortabbilder).

Bei den Experimenten zur Textdetektion wird die Detektionsrate berechnet. Er gibt die Prozentzahl an gefundenen Wörter des Datensatzes wieder. Hierbei liegt der Fo-

kus auf dem Lokalisieren der Wörter, ohne dabei auf die Erkennung der Wörter einzugehen. Zusätzlich zur Detektionrate wird die durchschnittliche Anzahl von Hypothesen je Seite angeben, die sich aus dem arithmetischen Mittel über alle Seiten des Datensatzes berechnet. Da nur die Dokumente des Krupp Datensatzes schwachen Kontrast aufweisen, werden nur diese vorverarbeitet (vgl. Abschnitt 4.1).

Für die Auswertung der segmentierungsbasierten Indizierung werden die WR und ∅WR (Unterabschnitt 5.2.3) verwendet. Hierbei soll mit der WR geprüft werden, wie viele Wörter richtig transkribiert wurden. Um einen Vergleich der Erkennungsrate über alle möglichen Annotationsaufwände zu geben, wird die ∅WR berechnet. Auch hier werden wieder beide Datensätze evaluiert. Durch Silbentrennung getrennte Wörter und grammatikalische Variationen von Wörtern werden jeweils als verschiedene Wörter gesehen. Für die Berechnung der ∅WR auf dem GW Datensatz wird das Verfahren mit 486, 972, 1458, 1944, 2430, 2916, 3402, 3888, 4374 Clustern ausgewertet. Der Krupp Datensatz wird aufgrund seines schwachen Kontrasts mit dem Histogrammausgleich und dem anschließenden Medianfilter vorverarbeitet. Bei dem Krupp Datensatz werden die Größen 76, 152, 228, 304, 380, 456, 532, 608, 684 zur Berechnung der ∅WR betrachtet.

Der segmentierungsfreie Fall wird auf auf dem GW Datensatz ausgewertet. Da die Durchführung der segmentierungsfreien Indizierung durch die hohe Anzahl an Hypothesen sehr zeitaufwändig ist, wird diese mit den besten ermittelten Konfigurationen evaluiert. Mit $Maß_1$ und $Maß_2$ (Unterabschnitt 5.2.4) werden zwei Metriken berechnet, die eine Bewertung der Qualität des Endergebnisses geben sollen.

## 5.4 AUSWERTUNG

Der folgende Abschnitt behandelt die durchgeführten Experimente dieser Arbeit. Zu Beginn wird mit den durchgeführten Experimenten eine Evaluierung der Parameter präsentiert, dessen Ziel es ist, die erzielten Ergebnisse weitestgehend zu optimieren. Dabei soll für die Textdetektion eine geeignete Konfiguration gefunden werden, mit der sowohl eine gute Detektionsrate als auch eine dazu im Verhältnis stehende Komplexität in der Anzahl der Worthypothesen erreicht wird (Unterabschnitt 5.4.1). Danach wird die segmentierungsbasierte Indizierung behandelt. Mit dem Ziel die Worterkennungsrate weiter zu steigern, wird in Unterabschnitt 5.4.2 eine Auswertung der Parameterkonfigurationen durchgeführt. Anschließend werden die Ergebnisse beider Teile zusammengeführt und die jeweils besten Werte für die Parameter zur Auswertung des segmentierungsfreien Fall gewählt. Am Ende findet ein Vergleich zum Stand

der Technik statt, um die Ergebnisse der hier verwendeten Methodik in Bezug mit anderen Verfahren zu setzen.

### 5.4.1 *Textdetektion*

Nach bestem Wissen und Gewissen des Autors ist der MSER Detektor in noch keiner Arbeit zur Textdetektion im Bereich der historischen handgeschriebenen Dokumente verwendet worden. Mit dem folgenden Abschnitt soll das Potential des MSER Detektor für diese Aufgabe gezeigt werden. Dafür werden zu Beginn die Einflüsse der einzustellenden Parameter auf die DR diskutiert. Für die Erläuterung der Parameter des MSER Detektors sei hier auf Kapitel 4 verwiesen. Als erstes werden die Ergebnisse der Parameterevaluierung auf dem GW Datensatz vorgestellt (Tabelle 5.4.1). Einen größeren Einfluss hat das gewählte $\Delta$. Bessere DR Werte werden mit kleineren $\Delta$ erreicht. Dies lässt sich inhärent dadurch erklären, dass mit kleinerem $\Delta$ die Bedingung der MSER Stabilität weniger streng ist, da Regionen so für weniger Binarisierungsschwellwerte stabil sein müssen. Bei zu hohen $\Delta$ werden mögliche Textstellen nicht detektiert. Mit $\Delta = 1$ wurde eine DR von $81,2\%$ erreicht. Die weiteren Experimente werden deswegen mit $\Delta = 1$ durchgeführt.

Als nächstes wird der Einfluss des Variationsparameters ausgewertet (Tabelle 5.4.2). Kleinere Werte für die maximale Variation bedeuten, dass MSER, welche weniger stabil sind, verworfen werden. Dadurch könnten potentielle Textstellen verworfen werden. Dies führt dazu, dass bei wachsender Variation die DR von $81,1\%$ auf $81,4\%$ gesteigert werden kann, wenn mehr MSER betrachtet werden. Die Anzahl der Hypothesen wächst verhältnismäßig wenig dazu.

In Tabelle 5.4.3 sind die Ergebnisse der Evaluierung des Parameters für den minimalen Flächenunterschied dargestellt. Bei kleineren Werten werden zwar weniger MSER verworfen, womit ein Anstieg der DR auf $84,9\%$ möglich ist, gleichzeitig steigt

| $\Delta$ | ∅Hypothesenzahl | DR |
|---|---|---|
| 5 | 49558 | 79,3% |
| 3 | 51619 | 80,5% |
| 2 | 51859 | 81,1% |
| 1 | 52164 | 81,2% |

Tabelle 5.4.1: GW: Einfluss von Delta auf die Detektionsrate und die durchschnittliche Hypothesenzahl.

| Maximale Variation | ∅Hypothesenzahl | DR |
|---|---|---|
| 0,20 | 52052 | 81,1% |
| 0,40 | 52340 | 81,4% |
| 0,60 | 52368 | 81,4% |
| 1,0 | 52338 | 81,4% |

Tabelle 5.4.2: GW: Einfluss des Parameters maximale Variation auf die DR und die durchschnittliche Hypothesenzahl.

| Minimaler Flächenunterschied | ∅Hypothesenzahl | DR |
|---|---|---|
| 0,2 | 52164 | 81,2% |
| 0,1 | 61055 | 84,6% |
| 0,05 | 61584 | 84,9% |
| 0,04 | 61533 | 84,9% |

Tabelle 5.4.3: GW: Einfluss des Parameters für den minimalen Flächenunterschied auf die DR und die durchschnittliche Hypothesenzahl.

aber die Anzahl an generierten Hypothesen stark. Da die Anzahl der Hypothesen zu sehr steigt, wird der Wert dieses Parameters auf $0,2$ festgelegt.

Im Weiteren werden verschiedene Werte für die Distanzwerte zur Hypothesengenerierung geprüft (Tabelle 5.4.4). Je größer diese sind, desto mehr Hypothesen werden erzeugt. Da im Weiteren keine Filterung von Worthypothesen durchgeführt wird, muss ein Kompromiss zwischen DR und der durchschnittlichen Hypothesenzahl getroffen werden, damit die Komplexität bei der weiteren Indizierung nicht zu groß wird. Distanzwerte für $d_x$ von 110-120 führten zu guten DR. Größere Distanzwerte werden nicht weiter betrachtet, da der Zuwachs an generierten Hypothesen zu keiner wesentlichen Verbesserung der DR führt. Der Distanzwert in y-Richtung war dabei auf 15 Pixel festgelegt. Kleinere Werte sind hierfür nötig, da sonst zu viele Hypothesen erzeugt werden, die Wörter unterschiedlicher Zeilen enthalten. In der letzten Zeile der Tabelle wurde der Distanzwert in y-Richtung von 15 Pixeln auf 20 erhöht, womit eine unmittelbare Steigerung der DR erzielt wurde, aber auch die Anzahl der Hypothesen stark steigt.

| $d_x$ | $d_y$ | ∅Hypothesenzahl | DR |
|---|---|---|---|
| 50 | 15 | 24889 | 79,9% |
| 80 | 15 | 39706 | 80,9% |
| 100 | 15 | 48138 | 81,1% |
| 110 | 15 | 52164 | 81,2% |
| 120 | 15 | 56098 | 81,3% |
| 130 | 15 | 59993 | 81,4% |
| 110 | 20 | 57026 | 83,5% |

Tabelle 5.4.4: GW: Einfluss der Distanzwerte auf die DR und die durchschnittliche Hypothesenzahl.

Auf dem Krupp Datensatz führt ein kleiner Wert für $\Delta$ zu einer besseren DR (Tabelle 5.4.5). Wird dieser auf 5 festgelegt , sinkt die DR. Es konnte ein Unterschied von bis 5% beobachtet werden.

Tabelle 5.4.6 präsentiert den Einfluss von verschiedenen Werten für die maximale Flächenvariation einer MSER über die $\Delta$ Schwellwerte. Auch hier hat das Erzeugen von mehr MSER einen positiven Einfluss auf die DR. Wird er eher klein gewählt mit $0,05$, sinkt die DR auf $71,8\%$. Wird die Bedingung weniger streng ausgelegt, kann die DR auf $86,7\%$ gesteigert werden.

Verschiedene Konfigurationen für den minimale Flächenunterschied werden in Tabelle 5.4.7 getestet. Größere Werte hierfür führen dazu, dass weniger MSER generiert werden. Dies hat zur Folge, dass weniger Hypothesen erzeugt werden. Dabei wird ein negativer Effekt auf die DR beobachtet. Das beste Ergebnis von $86,7\%$ wurde mit einem Wert von $0,04$ erreicht. Hierbei seien dennoch die Werte $0,20$ und $0.40$ hervorgehoben, mit denen zwar eine geringere DR erzielt wurde, die Hypothesenanzahl aber wesentlich kleiner ausfällt.

| $\Delta$ | ∅Hypothesenzahl | DR |
|---|---|---|
| 0 | 16589 | 84,5% |
| 1 | 15917 | 83,8% |
| 5 | 9414 | 79,5% |

Tabelle 5.4.5: Krupp: Einfluss von Delta auf die DR und die durchschnittliche Hypothesenzahl.

| Maximale Variation | ∅Hypothesenzahl | DR |
|---|---|---|
| 0,05 | 12994 | 71,8% |
| 0,10 | 26455 | 85,1% |
| 0,20 | 28740 | 86,2% |
| 0,30 | 28927 | 86,6% |
| 0,40 | 29081 | 86,6% |
| 0,80 | 29135 | 86,7% |

Tabelle 5.4.6: Krupp: Einfluss des Parameters maximale Variation auf die DR und die durchschnittliche Hypothesenzahl.

| Minimaler Flächenunterschied | ∅Hypothesenzahl | DR |
|---|---|---|
| 0,01 | 29131 | 86,7% |
| 0,04 | 28279 | 86,7% |
| 0,20 | 16324 | 86,5% |
| 0,40 | 9637 | 84,9% |

Tabelle 5.4.7: Krupp: Einfluss des Parameters für den minimalen Flächenunterschied auf die DR und die durchschnittliche Hypothesenzahl.

| $d_x$ | $d_y$ | ∅Hypothesenzahl | DR |
|---|---|---|---|
| 200 | 15 | 17366 | 83,6% |
| 200 | 20 | 25927 | 86,7% |
| 220 | 20 | 28287 | 86,7% |

Tabelle 5.4.8: Krupp: Einfluss der Distanzwerte auf die DR und die durchschnittliche Hypothesenzahl.

Das Verhalten von größeren Distanzwerten für das Gruppieren von MSER wird in Tabelle 5.4.8 gezeigt. Auch hier kann durch Erhöhen der Werte die DR weiter verbessert werden, mit dem Nachteil zusätzliche Hypothesen zu generieren. Auffällig ist der Anstieg von 3% bei der Erhöhung des Wertes für $d_y$, wobei beachtet werden muss, dass die Zahl der Hypothesen deutlich steigt. Hierbei sind aufgrund der nicht vorhandenen einheitlichen Schriftstruktur (Abbildung 5.1.2 (b)) der Dokumente größere Werte nötig.

Ein Problem bei der Auswertung der Textdetektion innerhalb des GW Datensatzes ist in Abbildung 5.4.1 dargestellt. Zu den Annotationen (rot) sind jeweils die Hypothesen mit dem besten IoU Wert abgebildet (grün). Dadurch, dass MSER Zusammenhangskomponenten von Schriftpixeln erfassen, entstehen oft präzise Hypothesen entlang der Wortkonturen. Wenn die Annotation unpräzise ist und weit über das Wort hinausgeht, dabei die Hypothese das Wort aber präzise abdeckt (vgl. Abbildung 5.4.1), fällt der IoU Wert unter 0.5. Obwohl das Wort mit der Hypothese korrekt detektiert wird, ist dies für die DR kein Treffer. Aus diesem Grund wird im Folgenden der Effekt von weniger strengen IoU Bedingungen ausgewertet. Eine weniger strenge Bedingung dessen, führt natürlich unmittelbar zu einer besseren DR, womit die Ergebnisse hierzu keinesfalls repräsentativ sind. Bei dem GW Datensatz (Tabelle 5.4.9) fällt aber auf, dass ein Heruntersetzen der Grenze des IoU Wert auf 0,4 eine erhebliche Verbesserung der DR (93, 6%) mit sich bringt. Es lässt sich somit vermuten, dass diese Fälle einen größeren Einfluss auf die DR haben.

Zusammenfassend kann geschlussfolgert werden, dass eine Textdetektion in historischen Dokumente mithilfe des MSER Detektor ein geeignetes und berechtigtes Vorgehen ist, um die Wörter eines Dokumentes zu segmentieren. Aufgrund der einfachen Parameterisierung und dem relativ konstantem Verhalten bei unterschiedlichen Werten für die Parameter, ist es möglich auf verschiedene Dokumentensammlungen zu abstrahieren. Somit setzt sich dieser gegenüber einer einfachen Binarisierung durch, da der Schwellwert für die Binarisierung für jede Dokumentensammlung neu bestimmt werden muss.

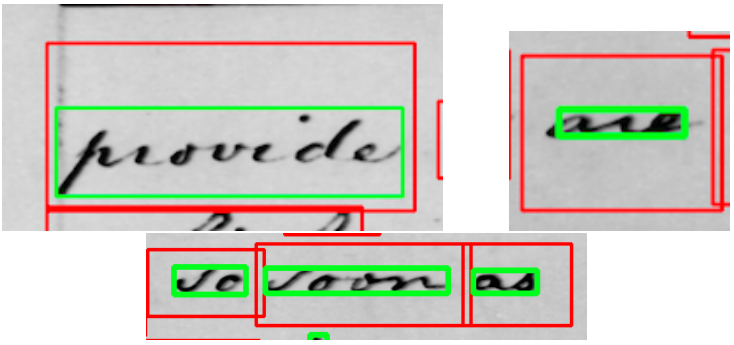


Abbildung 5.4.1: Ungenaue Annotationen erschweren die Textdetektion. Wenn die Hypothesen (grün) im Gegensatz zur Annotation (rot) zu präzise sind, wird das Wortabbild aufgrund des kleinen IOU Wert nicht als Treffer gewertet.

| Datensatz | IoU | DR |
|---|---|---|
| GW | 0,5 | 81,2% |
| GW | 0,4 | 93,6% |
| GW | 0,3 | 98,6% |
| Krupp | 0,5 | 83,9% |
| Krupp | 0,4 | 90,5% |
| Krupp | 0,3 | 95,2% |

Tabelle 5.4.9: DR mit verschieden strengen Bedingungen für die IoU.

### 5.4.2 *Segmentierungsbasierte Indizierung*

In diesem Abschnitt werden die Ergebnisse zur segmentierungsbasierten Indizierung vorgestellt. Tabelle 5.4.10 zeigt die Konsistenz und Reproduzierbarkeit des Verfahrens, trotz der zufälligen Initialisierung des Lloyd Algorithmus. Hierbei werden exemplarisch auf dem GW Datensatz für unterschiedliche Anzahlen k von Clustern jeweils fünf Wiederholungen der segmentierungsbasierten Indizierung ausgeführt. Zu jedem k wird das arithmetische Mittel der WR bestimmt (letzte Spalte) und die Standardabweichung angegeben (vorletzte Spalte). Mit größer werdendem k verringert sich die Standardabweichung von 1,11% auf unter 0,4%. Aufgrund der relativ kleinen Standardabweichung für in der Praxis relevante k (~1458) werden die weiteren Experimente nur ein einziges Mal durchgeführt.

Die Wahl der Anzahl k von Clustern hat einen großen Einfluss auf die Worterkennungsrate hat. Diese wächst mit der Anzahl der Cluster, womit aber der Annotationsaufwand steigt. Zudem ist es nicht sinnvoll, zu viele Cluster zu erstellen. Werden zu viele Cluster generiert, ist das Verfahren zu spezifisch und gleiche Wörter liegen in unterschiedlichen Clustern. Dies motiviert für die folgenden Experimente die ∅WR zu berechnen, die ein Mittel über alle möglichen k gibt.

Das Ziel der folgenden Experimente ist es, durch Konfigurieren der Parameter die Worterkennungsrate zu verbessern. In Tabelle 5.4.11 sind die Ergebnisse der Parameterevaluierung präsentiert. Für die SIFT Deskriptoren wird eine 4 x 4 Zellenunterteilung gewählt, die mit einer Schrittweite von 5 Pixeln in horizontaler und vertikaler Richtung berechnet werden (vgl. Abschnitt 2.2). Die Größe des visuellen Vokabulars der BoF (vgl. Abschnitt 2.3) orientiert sich an [RATL15]. Dort wurde für den GW gezeigt, dass eine Größe von 4096 sinnvoll ist und einen Kompromiss zwischen Dimensionalität und Abstrahierungsmöglichkeit liefert. Auch Rothacker et al. wählten dieselbe Größe in [RRF13].

Zu Beginn wird, wie in den ersten vier Zeilen der Tabelle zu sehen, die Größe der Zellen des SIFT Deskriptors angepasst. Mit einer Größe von 12 wurde eine ∅WR von 74,6% erreicht. Dies ergibt somit 48 x 48 Pixel große SIFT Deskriptoren. Sowohl größere (14) als auch kleinere Werte (8, 10) für die Zellengröße führten zu schlechteren Ergebnissen. Aus den Beobachtungen lässt sich folgern, dass zu kleine Deskriptoren nicht genügend Schriftverläufe beinhalten und somit nicht spezifisch genug sind. Ebenso fällt auf, dass bei zu großen Deskriptoren zu viele Pixel der unmittelbaren Umgebung betrachtet werden und somit zu spezifisch werden.

Für die Konfiguration der Spatial Pyramid Aufteilung wird der Wert 12 für die Zellengröße übernommen. Bei der Spatial Pyramid gilt es zu beachten, dass je feiner die Einteilungen der Spatial Pyramid gewählt werden, so größer die Dimensionalität der

| k | Standardabweichung | gemittelte WR |
|---|---|---|
| 486 | 1,11% | 56,44% |
| 972 | 0,53% | 66,38% |
| 1458 | 0,17% | 73,07% |
| 1944 | 0,33% | 78,62% |
| 2430 | 0,15% | 83,18% |
| 2916 | 0,36% | 87,02% |
| 3402 | 0,17% | 90,84% |
| 3888 | 0,12% | 94,08% |
| 4374 | 0,21% | 97,06% |

Tabelle 5.4.10: Einfluss der zufälligen Initialisierung des Lloyd Algorithmus.

Repräsentationen ausfällt. Dies muss im Weiteren bei der Wahl der Einteilung beachtet werden, da durch die Wahl der Vokabulargröße von 4096 die Merkmalsvektoren um ein Vielfaches davon wachsen. Mit feineren Einteilungen konnte die ∅WR weiter verbessert werden. Das beste Ergebnis von 75,2% ∅WR wurde mit einer Einteilung von 1x2 | 1x4 erreicht. Gröbere Einteilungen führten zu einem unspezifischerem Modell mit geringeren ∅WR. Bei feineren Einteilungen wurde das Modell zu spezifisch und führte zu geringeren ∅WR. Als beste Konfiguration hat sich somit die Zellengröße 12 und eine Spatial Pyramid Konfiguration von 1x2 | 1x4 ergeben.

Auf Basis dieser Konfigurationen wird eine reale Anwendung in der Praxis simuliert. Dafür ist die Zahl der Cluster notwendig. Eine Möglichkeit diese zu bestimmen wurde in [RM06] geliefert. Mit dem Gesetz von Heap ist die Zahl der verschiedenen Wörter auf 1365 geschätzt worden. Dieser Wert wird hier gewählt, um später einen Vergleich zum Stand der Technik geben zu können. Insgesamt konnte eine WR von 72,1% erzielt werden (letze Zeile).

Zu diesem Experiment erfolgt als nächstes eine qualitative Analyse. In Abbildung 5.4.2 sind berechnete Cluster und die Erkennungsrate innerhalb eines Clusters visualisiert. Grün eingerahmte Wortabbilder gelten als richtig und rot eingerahmte als falsch erkannt. Für den obersten Cluster wurde die Annotation „fort“ vergeben. Durch das Propagieren an die weiteren Wortabbilder werden alle Wortabbilder richtig transkribiert. Es fällt auf, dass das Wortabbild Fort sehr diskriminative Schriftverläufe besitzt, und somit ein fehlerfreier Cluster entsteht. Der mittlere Cluster wurde mit „receive“ annotiert. Auch hier ist zu sehen, dass der Cluster aus Wortabbildern besteht, die ähn-

| | k | Zellengröße | Spatial Pyramid | ∅WR | WR |
|---|---|---|---|---|---|
| | | 10 | 1x1 \| 1x2 | 74,2% | / |
| | | 8 | 1x1 \| 1x2 | 72,8% | / |
| | | 12 | 1x1 \| 1x2 | 74,6% | / |
| | | 14 | 1x1 \| 1x2 | 74,4% | / |
| | | 12 | 1x1 \| 1x4 | **75,1%** | / |
| | | 12 | 1x1 \| 1x6 | 74,7% | / |
| | | 12 | 1x2 \| 1x4 | **75,2%** | / |
| | | 12 | 1x2 \| 1x6 | 74,7% | / |
| | | 12 | 1x3 \| 1x6 | 74,8% | / |
| | | 12 | 1x4 \| 1x6 | 74,9% | / |
| | 1365 | 12 | 1x2 \| 1x4 | / | **72,1%** |

Tabelle 5.4.11: GW: Zellengröße der SIFT Deskriptoren und Spatial Pyramid Aufteilung. Bei der Berechnung für die ∅WR wurde die Indizierung mit 486, 972, 1458, 1944, 2430, 2916, 3402, 3888, 4374 Clustern durchgeführt.

liche Schriftverläufe aufweisen. Dies hat zur Folge, dass auch Wortabbilder, die zwar ähnlich aussehen, aber dennoch unterschiedlich sind, demselben Cluster zugewiesen werden. Angemerkt sei hier, dass einige Fehlerfälle grammatikalischer Natur sind (Vergangenheitsform des Wortes „receive“) und subjektiv nicht als Fehler einzuordnen sind. Da kein Stemming der Annotation vorgenommen wird, werden auch diese Fälle als falsch erkannte Wörter betrachtet. Der unterste Cluster wurde mit „company“ annotiert. Grammatikalische Variationen werden als fehlerhafte Transkription betrachtet. Das Wortabbild „compa“, welches durch Silbentrennung entstand, wird als falsche Zuweisung interpretiert. Die restlichen Fehlerfälle sind zu „company“ visuell ähnliche Wortabbilder.

Im Weiteren folgen die Ergebnisse auf dem Krupp Datensatz. Da der Krupp Datensatz manuell für diese Arbeit erstellt wurde, existieren zu diesem keine vergleichbaren Referenzwerte. Aus diesem Grund wird zunächst eine Evaluierung der Parameter durchgeführt und anschließend mit der besten Konfiguration ein Versuch ausgewertet, bei dem k gleich der Anzahl an verschiedenen Wörtern im Datensatz entspricht. Dadurch soll eine Anwendung in der Praxis simuliert werden (siehe Abschnitt 4.4 und 3.2.2). Für den Krupp Datensatz hat sich eine Zellengröße von 12 bewährt (Tabelle 5.4.12). Mit dieser konnte eine Steigerung der ∅WR um 1% beobachtet werden.

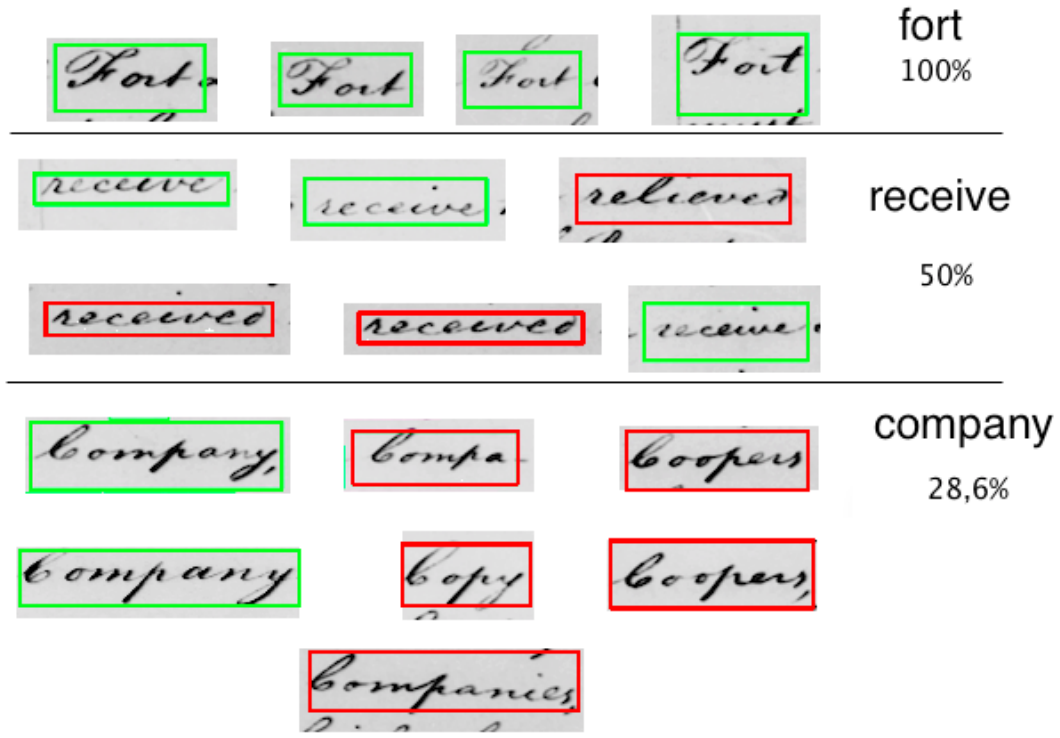



Abbildung 5.4.2: Qualitative Analyse von drei generierten Cluster. Grün eingerahmte Wortabbilder visualisieren korrekt erkannte Wörter, während rote eingerahmte Wortabbilder fehlerhafte Transkriptionen bedeuten. Jeweils die Annotationen der Cluster und die clusterspezifische Erkennungsraten sind angegeben.

Als beste Spatial Pyramid Einteilung hat sich 1x2 | 1x6 herausgestellt mit einer ∅WR von insgesamt 53,5%.

Mit diesen Einstellungen wird ein praxisbezogenes Experiment ausgeführt, mit dem Ziel ein Anwendung in der Realität zu simulieren (letzte Zeile). Als $k$ wird dabei 446 festgelegt, die Zahl der verschiedenen Wörter. In der Praxis ist ein Wert dafür aber nicht bekannt und müsste geschätzt werden. Für das Experiment ist die Zahl mithilfe der Annotation bestimmt worden. Hier sei auf Unterabschnitt 3.2.2 verwiesen, in dem der Wert über das Gesetz von Heap approximiert wurde. Mit der hier vorgeschlagenen Methode konnte eine WR von 62,4% erreicht werden. Die geringere WR ist durch die Schwierigkeit des Datensatzes zu erklären (Abschnitt 5.1). Der SIFT Deskriptor scheint für das Schriftbild des Datensatzes nicht charakteristisch genug

| | k | Zellengröße | Spatial Pyramid | ∅WR | WR |
|---|---|---|---|---|---|
| | | 8 | 1x1 \| 1x8 | 52,2% | / |
| | | 10 | 1x1 \| 1x8 | 52,7% | / |
| | | 12 | 1x1 \| 1x8 | 53,2% | / |
| | | 12 | 1x1 \| 1x2 | 52,6% | / |
| | | 12 | 1x2 \| 1x6 | **53,5%** | / |
| | 446 | 12 | 1x2 \| 1x6 | / | **62,4%** |

Tabelle 5.4.12: Krupp: Zellengröße der SIFT Deskriptoren und Spatial Pyramid Aufteilung. Bei der Berechnung für die ∅WR wurde die Indizierung mit 76,152,228,304,380,456,532,608,684 Clustern durchgeführt.

zu sein und somit können die verschiedenen Schriftverläufen nicht weit genug unterschieden werden. Da die Wortabmessungen innerhalb des Datensatzes stark variieren können, ist auch die feste Größe der Deskriptoren (Zellengröße) ein Faktor, der die WR negativ beeinflusst.

### 5.4.3 *Segmentierungsfreie Indizierung*

In diesem Teil wird die Bewertung der segmentierungsfreien Indizierung präsentiert, die bislang in noch keiner Arbeit behandelt wurde. Die Ergebnisse dieses Abschnittes sind somit nach bestem Wissen und Gewissen des Autors eine Neuheit für die Dokumentenanalyse. Da die große Anzahl an Hypothesen zu einer hohen Laufzeit bei der Auswertung führt, ist eine weitere ausführliche Evaluierung der Parameter für den segmentierungsfreien Fall nicht möglich. Aus diesem Grund werden die in den vorherigen Teilen bestimmten Konfigurationen übernommen. Um auch hier eine realitätsbezogene Anwendung zu simulieren, wird im Weiteren das k festgelegt. Für den GW Datensatz sind dies 1365 Cluster. Die Ergebnisse hierfür sind in Tabelle 5.4.13 aufgeführt. Für das $Maß_1$ konnte ein Wert von 35,6% erzielt werden. Für die strengere Metrik $Maß_2$ erreichte die vorgeschlagene Methodik 24,2%. Die geringen Werte lassen sich durch die hohe Anzahl an Hypothesen begründen. Dadurch, dass keine explizite Filterung von Worthypothesen durchgeführt wird, werden viele fehlerhafte Segmente geclustert, womit die schlechteren Werte zu erklären sind.

In Abbildung 5.4.3 sind drei von dem Verfahren generierte Cluster visualisiert. Je Cluster sind jeweils die fünf nächsten Nachbarn der Clusterzentroiden dargestellt. Das obersten Cluster wurde mit „immediately“ annotiert. Hierbei entstand kein sinn-

voller Cluster. Für den mittleren Cluster wurde die Annotation „until“ vergeben. Die Worthypothesen umfassen ähnliche Schriftverläufe. Es ist eine Hypothese vorhanden, die das Wortabbild optimal einrahmt und korrekt erkannt wurde. Auffällig ist der unterste Cluster, der mit „december“ annotiert wurde. Die abgebildeten Hypothesen des Clusters besitzen ähnliche Schriftverläufe. Es handelt sich bei den Hypothesen jedoch um keine holistischen Wortabbilder und es entsteht ein qualitativ schlechter Cluster.

Beim segmentierungsfreien QbE Word Spotting können unpassende Wortkandidaten durch die Abmessung der von Nutzer gestellten Anfrage verworfen werden. Zusätzlich gilt es nur die visuell ähnlichsten Worthypothesen zu finden, wodurch es über non-maximum-suppression möglich ist, schlechtere, überlappende Wortkandidaten zu unterdrücken [RRF13]. Im Gegensatz hierzu müssen alle generierten Worthypothesen betrachtet werden. Dies hat somit einen starken Einfluss auf die Zentroidbedingung, womit die schlechteren Werte zu erklären sind.

| | k | Maß$_1$ | Maß$_2$ |
|---|---|---|---|
| Vorgeschlagene Methode | 1365 | 35,6% | 24,2% |

Tabelle 5.4.13: GW: Ergebnisse der segmentierungsfreien Indizierung.


Abbildung 5.4.3: Qualitative Analyse der segmentierungsfreien Indizierung anhand von drei generierten Cluster. Grün eingerahmt sind die Worthypothesen.

## 5.5 VERGLEICH ZUM STAND DER TECHNIK

Im Folgenden sollen die erreichten Ergebnisse der Arbeit mit dem Stand der Technik in Relation gesetzt werden. Direkt vergleichbar sind die Ergebnisse aus [RM06]. Die Arbeit befasst sich mit der segmentierungsbasierten Indizierung und wurde in Unterabschnitt 3.2.2 beschrieben. Zur Evaluierung wurde derselbe GW Datensatz benutzt. Aus diesem Grund wird die hier vorgeschlagene Methode mit der von den Autoren bestimmten Anzahl von Clustern ausgewertet. Diese wurde für den Datensatz auf 1365 geschätzt. Mit der besten Konfiguration und der Anzahl an zu erstellenden Cluster wurde mit der hier vorgeschlagenen Methodik eine WR von 72,1% erzielt (siehe Tabelle 5.5.1 3. Zeile). Dieser Wert übertrifft den in [RM06] erreichten Stand der Technik von 68,5%. Es sei zudem angemerkt, dass das dort benutzte Evaluierungsprotokoll nicht wie hier vorsah, die Cluster über die Medoiden zu annotieren. Stattdessen werden die Cluster mit dem am häufigsten vorkommendem Wort inner-

halb des Clusters transkribiert. Dies bedeutet insbesondere, dass dabei eine größere Anzahl an Wortabbildern betrachtet werden müssen. Zwar gilt es nur die 1365 Cluster zu annotieren, dennoch wird hier vom Nutzer ein erheblicher Aufwand gefordert. Dieser wird im weiteren als Beobachtungsaufwand bezeichnet. Ein Experiment, welches denselben zusätzlichen Beobachtungsaufwand verwendete, ist in der 2. Zeile aufgeführt (72,5%). Die erzielten WR für beide Annotationsmöglichkeiten unterscheiden sich nur geringfügig. Es lässt sich also festhalten, dass die Modellierung eines Indexeintrages über den Medoiden eine geeignete Wahl ist, die sowohl einen viel geringeren Beobachtungsaufwand mit sich bringt als auch bessere Ergebnisse im Vergleich zu [RM06] erzielen. Als wesentliche Begründung der besseren Ergebnisse ist die hier verwendete Merkmalsrepräsentation (BoF+Spatial Pyramid), die keine Wortnormalisierung benötigt, zu sehen. Die BoF Darstellung mit dem gradientenbasierten SIFT Deskriptor eignet besonders gut zur Modellierung von Wortabbildern.

Mit der Indizierung ist es möglich Dokumentensammlungen zu transkribieren. Somit sind auch die Ergebnisse aus [LRM04] mit der erzielten WR vergleichbar. Dort wurde ein HMM zur Handschrifterkennung trainiert, welches eine größere Menge an annotierten Trainingsdaten benötigt (vgl. Unterabschnitt 3.2.3). Für den GW Datensatz ist keine explizite Einteilung in Test- und Trainingsdaten definiert. Ein Protokoll, welches dennoch eine Güte über die Erkennungsrate auf neuen Dokumenten gibt, die nicht für das Lernen benutzt wurden, ist die Kreuzvalidierung. Dafür wird der Datensatz in der Arbeit jeweils in 19 Trainings- und 1 Testseiten aufgeteilt, sodass jede Seite einmal separat als Testseite ausgewertet wird, während die restlichen 19 Seiten für das Lernen des Modells verwendet werden. In jedem Durchlauf wurde auf der Testseite eine WR berechnet. Durch das Mitteln über alle 20 Durchläufe ergab sich eine mittlere WR von 55,1% (4. Tabellenzeile). Dasselbe Evaluierungsprotokoll wurde auch in [HRM05] verwendet. Mit einer Kreuzvalidierung auf dem GW Daten-

| | Annotationsaufwand | Beobachtungsaufwand | WR |
|---|---|---|---|
| Rath, Manmatha | 1365 | 4860 | 68,5% |
| **vorgeschlagene Methode** | 1365 | 4860 | **72,5%** |
| **vorgeschlagene Methode** | 1365 | 1365 | **72,1%** |
| Lavrenko et al. | ∅4617 | / | ∅55,1% |
| Howe et al. | ∅4617 | / | ∅63,6% |

Tabelle 5.5.1: Vergleich mit dem Stand der Technik. Alle Arbeiten funktionieren segmentierungsbasiert.

satz wurde eine mittlere WR von 63,6% erreicht (letzte Zeile). Sowohl die 55,1% aus [LRM04] als auch die 63,6% von [HRM05] konnten mit einem Annotationsaufwand von 1365 Wortabbildern übertroffen werden. Im Vergleich dazu befanden sich auf 19 Trainingsseiten durchschnittlich 4617 Wörter, die es vorher zu annotieren galt. Mit einem $\sim\frac{1}{3}$ an Annotationsaufwand konnte die Worterkennungsrate beider Arbeiten überboten werden. Grundsätzlich muss aber erwähnt werden, dass die automatische Transkription über Handschrifterkennung weitaus schwieriger ist, da ein Modell gelernt werden muss, welches anhand des visuellen Aussehens eines Wortabbildes eine Entscheidung über die zu vergebende Transkription treffen muss. Zudem ist das QbE Word Spotting inhärent einfacher, weil dieses sich „nur“ mit der visuellen Ähnlichkeitsbewertung von Wortabbildern befasst. An dieser Stelle sei hier auf die wesentlichen Unterschiede, die in Kapitel 3 ausgeführt sind, verwiesen.

# 6 FAZIT

Die automatische Transkription von historischen Dokumenten ist eine schwierige Aufgabe innerhalb der Dokumentenanalyse. Ein Problem bei der Klassifikation von Handschrift ist die Variabilität im visuellen Aussehen, die selbst bei gleichen Wörtern hoch sein kann. Holistische Worterkenner bieten hierfür ein Verfahren. Diese benötigen jedoch eine größere Mengen an Trainingsdaten.

Eine alternative Methode zur Generierung einer Transkription ist die Indizierung. Die grundlegende Idee der Indizierung ist es visuell ähnliche Wörter zu gruppieren. Es finden sich somit starke Parallelen zum QbE Word Spotting. Aus diesem Grund wurde in dieser Arbeit eine Merkmalsdarstellung verwendet, die sich im Bereich des QbE Word Spotting bewährt hat (siehe Kapitel 3). Mit dem SIFT Deskriptor, als hier verwendete Grundlage einer Bag of Features Darstellung und der darauf aufbauenden Spatial Pyramid Repräsentation, ist eine geeignete Merkmalsrepräsentation für Wortabbilder gegeben. Durch die einfache Darstellung als Histogramme gleicher Dimensionen kann so auf komplexe Verfahren, wie das HMM und die DTW, verzichtet werden. Für die Gruppierung der Wortabbilder wurde hier ein Lloyd Algorithmus auf Basis eines winkelbasierten Maßes benutzt, der bisher nach bestem Wissen und Gewissen des Autors in dieser Form für diese Aufgabe noch nicht benutzt wurde.

Damit wurde die Worterkennungsrate des aktuellen Standes der Technik auf dem GW Benchmark aus [RM06], welches auch die Indizierung behandelt, übertroffen (72,1%). Ein weiterer Vorteil, der sich gegenüber dem dort verwendeten Verfahren ergeben hat, ist der wesentlich geringere Beobachtungsaufwand (siehe Kapitel 5). Anstatt ein Cluster über das am häufigsten vorkommenden Wort zu annotieren, reicht die Betrachtung des repräsentativen Medoiden eines Clusters. Damit konnte der zusätzliche Aufwand bei der Annotation deutlich verringert werden. Ein Nachteil des Lloyd Algorithmus gegenüber den dort verwendeten agglomerativen Algorithmen ist die zufällige Initialisierung. Es konnte gezeigt werden, dass dieser Nachteil nur eine untergeordnete Rolle hat.

Ebenso konnte sich der Ansatz gegenüber holistischen Worterkennern durchsetzen (siehe Kapitel 5). Da die Worterkennungsrate des Verfahrens von der Anzahl der Cluster und somit dem Annotationsaufwand abhängig ist, wurde für diese Arbeit ein neues Evaluierungsmaß für die Indizierung vorgestellt. Mit der ∅WR konnte die Worterkennungsrate der Indizierung über unterschiedlich großen Annotationsaufwand

optimiert werden. Dabei konnten die Worterkennungsraten dieser Verfahren mit einem wesentlich geringeren Annotationsaufwand übertroffen werden. Aufgrund der beobachteten Ergebnisse lässt sich also folgern, dass die semi-überwachte Indizierung und der damit verbundenen nächsten Nachbar Klassifikation im Clustering Verfahren eine geeignetere Wahl zur Wortklassifikation ist.

Für die Arbeit wurde mit dem Krupp Datensatz ein neuer Benchmark erstellt. Die Qualität der Dokumente des Krupp Datensatzes ist durch das Altern der Dokumente stark beeinträchtigt. Zusätzlich mit der verwendeten Schriftart Kurrent und der unbeschränkten Schriftform ist ein schwieriger Datensatz angefertigt worden. Wegen dieser Schwierigkeiten wurde auf diesem Datensatz eine geringere Worterkennungsrate (62,4%) erzielt.

Mit der segmentierungsfreien Indizierung wurde ein Aufgabenbereich vorgestellt, der nach bestem Wissen und Gewissen des Autors zum ersten Mal behandelt worden ist. In Kapitel 3 wurde die Problematik der Segmentierung innerhalb historischer Dokumente diskutiert. Eine optimale Segmentierung innerhalb historischer Dokumente ist aufgrund der unbeschränkten Struktur von Handschrift nicht möglich, da die Segmentierung und Erkennung gekoppelt sind. Für ein optimales Segment müsste erkannt werden, ob es sich bei diesem um ein holistisches Wortabbild handelt, während im Gegenzug für die Erkennung ein optimales Segment benötigt wird. Patch-basierte Ansätze umgehen diesem Problem, indem ein Dokument in regelmäßige Bildausschnitte unterteilt wird [RATL15, RATL11, RFB+13]. Bei der Indizierung ist eine Größenanpassung der Bildausschnitte durch eine Anfrage nicht möglich. Aus diesem Grund wurde hier eine hypothesenbasierte Textdetektion umgesetzt. Grundlage dessen war der MSER Detektor, welcher nach bestem Wissen und Gewissen des Autors hier zum ersten Mal innerhalb historischer Dokumente verwendet worden ist. Insgesamt lieferte dieser Ansatz auf dem GW Benchmark akzeptable Detektionsergebnisse (~80%), in denen ein Großteil der nicht erkannten Wörter auf ungenaue Annotationen zurückzuführen sind. Auf dem Krupp Datensatz konnte eine wesentliche bessere Detektionsrate von 86,7% erreicht werden. Es lässt sich dennoch festhalten, dass der MSER Detektor ein geeignetes Mittel zur Textdetektion in historischen Dokumenten liefert.

Die Arbeit soll als Schritt in Richtung segmentierungsfreier Indizierung dienen. Bei der hier verwendeten Methode wird davon ausgegangen, dass jede Hypothese ein potentielles Wortabbild ist. Es wird keine weitere Aussortierung von Hypothesen, die keinen holistischen Wörtern entsprechen, durchgeführt. Diese hohe Anzahl an fehlerhaften Hypothesen hat aber einen direkten Einfluss auf das Clustering. Beim segmentierungsfreien Word Spotting kann angenommen werden, dass eine Anfrage vom Nutzer gestellt wird, die mit einem optimal segmentierten Wortabbild korrespon-

diert. Mit dieser Anfrage können relevante Bildausschnitte gefunden werden. Durch non-maximum-suppression ist es hierbei möglich, nur die zur Anfrage passendsten Bildausschnitte zu extrahieren und dabei schlechtere, überlappende Worthypothesen zu unterdrücken [RRF13]. Im Gegensatz hierzu ist bei der segmentierungsfreien Indizierung keine konkrete Anfrage gegeben, mit der ohne Weiteres Hypothesen verworfen werden können. Eine Idee mit der das Verfahren erweiterbar wäre, ist das Einbinden von Nutzerrückmeldungen. Nach dem initialen Clustering könnte hier statt dem direkten Annotieren der Cluster, ein Nutzer fehlgeschlagene Cluster, die aus schlechten Worthypothesen bestehen, entfernen. Mit den restlichen Hypothesen könnte das Verfahren erneut wiederholt werden, um somit schlechte Worthypothesen zu unterdrücken.

Insgesamt lässt sich dennoch festhalten, dass die vorgestellte Methode ein geeignetes Verfahren zur Transkription von historischen Dokumenten darstellt. Existiert eine gute Segmentierung, können damit sogar die Ergebnisse anderer segmentierungsbasierter Verfahren übertroffen werden. Aus den Beobachtungen zum segmentierungsbasierten Fall folgt, dass ein weiterer Klassifikator, der fehlerhafte Hypothesen verwirft, unmittelbar zu einer Leistungssteigerung führen würde.

# LITERATURVERZEICHNIS


[AGFV12] Almazán, Jon ; Gordo, Albert ; Fornés, Alicia ; Valveny, Ernest: Efficient Exemplar Word Spotting. In: *British Machine Vision Conference, BMVC 2012, Surrey, UK, September 3-7, 2012*, 2012, 1–11

[AR13] Aggarwal, Charu C. ; Reddy, Chandan K.: *Data Clustering: Algorithms and Applications*. 1st. Chapman & Hall/CRC, 2013. – ISBN 1466558210, 9781466558212

[ASSM10] Albalate, Amparo ; Suendermann, David ; Suchindranath, Aparna ; Minker, Wolfgang: A Semi-Supervised Cluster-and-Label Approach for Utterance Classification. In: *International Conference on Speech and Language Processing (Interspeech)*, 2010

[BDGS05] Banerjee, A. ; Dhillon, I. ; Ghosh, J. ; Sra, S.: Clustering on the Unit Hypersphere using von Mises-Fisher Distributions. In: *Journal of Machine Learning Research* 6 (2005), September, S. 1345–1382

[Bor14] Borovikov, Eugene: A survey of modern optical character recognition techniques. In: *CoRR* abs/1412.4183 (2014). http://arxiv.org/abs/1412.4183

[Bra00] Bradski, G.: In: *Dr. Dobb's Journal of Software Tools* (2000). http://opencv.org

[Bur98] Burges, Chris J.: A Tutorial on Support Vector Machines for Pattern Recognition, 1998, 121-167

[BZMn07] Bosch, Anna ; Zisserman, Andrew ; Muñoz, Xavier: Image Classification using Random Forests and Ferns. In: *ICCV*, IEEE, 2007, 1-8

[CCC+11] Coates, A. ; Carpenter, B. ; Case, C. ; Satheesh, S. ; Suresh, B. ; Wang, Tao ; Wu, D.J. ; Ng, A.Y.: Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning. In: *International Conference on Document Analysis and Recognition (ICDAR)*, 2011. – ISSN 1520–5363, 440–445

[CTS+11] Chen, Huizhong ; Tsai, Sam S. ; Schroth, Georg ; Chen, David M. ; Grzeszczuk, Radek ; Girod, Bernd: Robust Text Detection in Natural Images with Edge-enhanced Maximally Stable Extremal Regions. In: *2011 IEEE International Conference on Image Processing*. Brussels, September 2011

[DDF+90] Deerwester, Scott ; Dumais, Susan T. ; Furnas, George W. ; Landauer, Thomas K. ; Harshman, Richard: Indexing by latent semantic analysis. In: *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE* 41 (1990), Nr. 6, S. 391–407

[DM01] Dhillon, Inderjit S. ; Modha, Dharmendra S.: Concept Decompositions for Large Sparse Text Data Using Clustering. In: *Mach. Learn.* 42 (2001), Januar, Nr. 1-2, 143–175. http://dx.doi.org/10.1023/A:1007612920971. – DOI 10.1023/A:1007612920971. – ISSN 0885–6125

[DNLP16] Dey, Sounak ; Nicolaou, Anguelos ; Lladós, Josep ; Pal, Umapada: Evaluation of the Effect of Improper Segmentation on Word Spotting. In: *CoRR* abs/1604.06243 (2016). http://arxiv.org/abs/1604.06243

[EOW10] Epshtein, Boris ; Ofek, Eyal ; Wexler, Yonatan: Detecting text in natural scenes with stroke width transform. In: *CVPR*, IEEE, 2010, 2963-2970

[Fin14] Fink, Gernot A.: *Markov Models for Pattern Recognition, From Theory to Applications*. 2. London : Springer, 2014 (Advances in Computer Vision and Pattern Recognition)

[FR14] Fink, Gernot ; Rothacker, Leonard: *Statistic Models for Handwriting Recognition and Retrieval*. 2014

[FRG14] Fink, Gernot A. ; Rothacker, Leonard ; Grzeszick, Rene: Grouping Historical Postcards Using Query-by-Example Word Spotting. In: *Proc. Int. Conf. on Frontiers in Handwriting Recognition*. Crete, Greece, 2014

[FS96] Freund, Yoav ; Schapire, Robert E.: *Experiments with a New Boosting Algorithm*. 1996

[GW06] Gonzalez, Rafael C. ; Woods, Richard E.: *Digital Image Processing (3rd Edition)*. Upper Saddle River, NJ, USA : Prentice-Hall, Inc., 2006. – ISBN 013168728X

[HFKB12] Hornik, Kurt ; Feinerer, Ingo ; Kober, Martin ; Buchta, Christian: Spherical k-means clustering. In: *Journal of Statistical Software* 50 (2012), Nr. 10, S. 1–22

[HRM05] Howe, Nicholas R. ; Rath, Toni M. ; Manmatha, R.: Boosted Decision Trees for Word Recognition in Handwritten Document Retrieval. In: *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA : ACM, 2005 (SIGIR '05). – ISBN 1–59593–034–5, 377–383

[HTF01] Hastie, Trevor ; Tibshirani, Robert ; Friedman, Jerome: *The Elements of Statistical Learning*. New York, NY, USA : Springer New York Inc., 2001 (Springer Series in Statistics)

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

[KWD14] Kovalchuk, A. ; Wolf, L. ; Dershowitz, N.: A Simple and Fast Word Spotting Method. In: *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, 2014. – ISSN 2167–6445, S. 3–8

[LKA16] Lever, Jake ; Krzywinski, Martin ; Altman, Naomi: Points of Significance: Model selection and overfitting. In: *Nat Meth* 13 (2016), 09, Nr. 9, 703–704. http://dx.doi.org/10.1038/nmeth.3968. ISBN 1548–7091

[Low04] Lowe, David G.: Distinctive Image Features from Scale-Invariant Keypoints. In: *International Journal of Computer Vision* 60 (2004), Nr. 2, 91–110. http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94. – DOI 10.1023/B:VISI.0000029664.99615.94. – ISSN 1573–1405

[LRM04] Lavrenko, Victor ; Rath, Toni ; Manmatha, R.: Holistic Word Recognition for Handwritten Historical Documents. In: *Proceedings of Document Image Analysis for Libraries* PARC, Institute of Electrical & Electronics Engineering, January 2004, S. 278–287

[LSP06] Lazebnik, S. ; Schmid, C. ; Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* Bd. 2, 2006. – ISSN 1063–6919, S. 2169–2178

[Luh58] Luhn, H. P.: The Automatic Creation of Literature Abstracts. In: *IBM J. Res. Dev.* 2 (1958), April, Nr. 2, 159–165. http://dx.doi.org/10.1147/rd.22.0159. – DOI 10.1147/rd.22.0159. – ISSN 0018–8646

[MCUP02] Matas, J. ; Chum, O. ; Urban, M. ; Pajdla, T.: Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. In: *Proceedings of the British Machine Vision Conference*, BMVA Press, 2002. – ISBN 1–901725–19–7, S. 36.1–36.10. – doi:10.5244/C.16.36

[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. – ISSN 1063–6919, S. 631–637

[ML11] Murtagh, Fionn ; Legendre, Pierre: Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm. In: *CoRR* abs/1111.6285 (2011). http://dblp.uni-trier.de/db/journals/corr/corr1111.html#abs-1111-6285

[MR05] Manmatha, R. ; Rothfeder, Jamie L.: A Scale Space Approach for Automatically Segmenting Words from Historical Handwritten Documents. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 27 (2005), Nr. 8, 1212-1225. http://dblp.uni-trier.de/db/journals/pami/pami27.html#ManmathaR05

[MTS+05] Mikolajczyk, K. ; Tuytelaars, T. ; Schmid, C. ; Zisserman, A. ; Matas, J. ; Schaffalitzky, F. ; Kadir, T. ; Gool, L. V.: A Comparison of Affine Region Detectors. In: *Int. J. Comput. Vision* 65 (2005), November, Nr. 1-2, 43–72. http://dx.doi.org/10.1007/s11263-005-3848-x. – DOI 10.1007/s11263–005–3848–x. – ISSN 0920–5691

[NFHS11] Nischwitz, A. ; Fischer, M. ; Haberäcker, P ; Socher, G.: *Computergrafik und Bildverarbeitung*. Bd. 3. Auflage Band 2. Wiesbaden : Vieweg +Teubner Verlag | Springer Fachmedien Wiesbaden GmbH 2011, 2011

[NM65] Nelder, J. A. ; Mead, R.: A Simplex Method for Function Minimization. In: *Computer Journal* 7 (1965), S. 308–313

[NM11a] Neumann, L. ; Matas, J.: Text Localization in Real-World Images Using Efficiently Pruned Exhaustive Search. In: *2011 International Conference on Document Analysis and Recognition*, 2011. – ISSN 1520–5363, S. 687–691

[NM11b] Neumann, Lukas ; Matas, Jiri: A Method for Text Localization and Recognition in Real-world Images. In: *Proceedings of the 10th Asian Conference on Computer Vision - Volume Part III*. Berlin, Heidelberg : Springer-Verlag, 2011 (ACCV'10). – ISBN 978–3–642–19317–0, 770–783

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

[PD08] Pomberger, G. ; Dobler, H.: *Algorithmen und Datenstrukturen: eine systematische Einführung in die Programmierung*. Pearson Studium, 2008 (Pearson Studium - IT). https://books.google.de/books?id=SVtnGzSmGuUC. – ISBN 9783827372680

[Pia14] Piantadosi, Steven T.: Zipf's word frequency law in natural language: A critical review and future directions. In: *Psychonomic Bulletin & Review* 21 (2014), Nr. 5, 1112–1130. http://dx.doi.org/10.3758/s13423-014-0585-6. – DOI 10.3758/s13423–014–0585–6. – ISSN 1531–5320

[PKG11] Petitjean, François ; Ketterlin, Alain ; Gançarski, Pierre: A Global Averaging Method for Dynamic Time Warping, with Applications to Clustering. In: *Pattern Recogn.* 44 (2011), März, Nr. 3, 678–693. http://dx.doi.org/10.1016/j.patcog.2010.09.013. – DOI 10.1016/j.patcog.2010.09.013. – ISSN 0031–3203

[QM16] Qin, Siyang ; Manduchi, Roberto: A Fast and Robust Text Spotter. In: *IEEE Winter Conference on Computer Vision (WACV 2016)*, 2016

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

[RATL15] Rusiñol, Marçal ; Aldavert, David ; Toledo, Ricardo ; Lladós, Josep: Efficient segmentation-free keyword spotting in historical document collections. In: *Pattern Recognition* 48 (2015), Nr. 2, 545 - 555. http://dx.doi.org/http://dx.doi.org/10.1016/j.patcog.2014.08.021. – DOI http://dx.doi.org/10.1016/j.patcog.2014.08.021. – ISSN 0031–3203

[RF15] Rothacker, Leonard ; Fink, Gernot A.: Segmentation-free Query-by-String Word Spotting with Bag-of-Features HMMs. In: *Proc. Int. Conf. on Document Analysis and Recognition*. Nancy, France, 2015

[RF16] Rothacker ; Fink: *Fachprojekt Dokumentenanalyse*. 2016

[RFB+13] Rothacker, Leonard ; Fink, Gernot A. ; Banerjee, Purnendu ; Bhattacharya, Ujjwal ; Chaudhuri, Bidyut B.: Bag-of-Features HMMs for Segmentation-Free Bangla Word Spotting. In: *International Workshop on Multilingual OCR (MOCR)*. Washington DC, USA, 2013

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

[RM06] Rath, Tony M. ; Manmatha, R.: Word spotting for historical documents. In: *International Journal of Document Analysis and Recognition (IJDAR)* 9 (2006), Nr. 2, 139–152. http://dx.doi.org/10.1007/s10032-006-0027-8. – DOI 10.1007/s10032–006–0027–8. – ISSN 1433–2825

[RP08] Rodríguez, Jose A. ; Perronnin, Florent: Local gradient histogram features for word spotting in unconstrained handwritten documents. In: *Proceedings of the 1st International Conference on Handwriting Recognition (ICFHR'08)*, 2008

[RRF13] Rothacker, L. ; Rusiñol, M. ; Fink, G. A.: Bag-of-Features HMMs for Segmentation-Free Word Spotting in Handwritten Documents. In: *2013 12th International Conference on Document Analysis and Recognition*, 2013. – ISSN 1520–5363, S. 1305–1309

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

[RSP09] Rodríguez-Serrano, José A. ; Perronnin, Florent: Handwritten Word-spotting Using Hidden Markov Models and Universal Vocabularies. In: *Pattern Recogn.* 42 (2009), September, Nr. 9, 2106–2116. http://dx.doi.org/10.1016/j.patcog.2009.02.005. – DOI 10.1016/j.patcog.2009.02.005. – ISSN 0031–3203

[RVF12] Rothacker, L. ; Vajda, S. ; Fink, G. A.: Bag-of-Features Representations for Offline Handwriting Recognition Applied to Arabic Script. In: *Fron-*

*tiers in Handwriting Recognition (ICFHR), 2012 International Conference on*, 2012, S. 149–154

[RVGF14] Richarz, Jan ; Vajda, Szilard ; Grzeszick, Rene ; Fink, Gernot A.: Semi-Supervised Learning for Character Recognition in Historical Archive Documents. In: *Pattern Recognition, Special Issue on Handwriting Recognition* 47 (2014), Nr. 3, S. 1011 – 1020

[SC+12] Shrivakshan, GT ; Chandrasekar, C u. a.: A comparison of various edge detection techniques used in image processing. In: *IJCSI International Journal of Computer Science Issues* 9 (2012), Nr. 5, S. 272–276

[SL91] Safavian, S. R. ; Landgrebe, D.: A survey of decision tree classifier methodology. In: *IEEE Transactions on Systems, Man, and Cybernetics* 21 (1991), May, Nr. 3, S. 660–674. http://dx.doi.org/10.1109/21.97458. – DOI 10.1109/21.97458. – ISSN 0018–9472

[SNL14] Shi, R. ; Ngan, K. N. ; Li, S.: Jaccard index compensation for object segmentation evaluation. In: *2014 IEEE International Conference on Image Processing (ICIP)*, 2014. – ISSN 1522–4880, S. 4457–4461

[Ste03] Steland, A.: *Mathematische Grundlagen Der Empirischen Forschung*. Springer Berlin Heidelberg, 2003 (Statistik und ihre Anwendungen). https://books.google.de/books?id=aTws5c57pt4C. – ISBN 9783540037002

[TD15] Tabassum, A. ; Dhondse, S. A.: Text Detection Using MSER and Stroke Width Transform. In: *Communication Systems and Network Technologies (CSNT), 2015 Fifth International Conference on*, 2015, S. 568–571

[WBB11] Wang, Kai ; Babenko, Boris ; Belongie, Serge: End-to-End Scene Text Recognition. In: *IEEE International Conference on Computer Vision (ICCV)*. Barcelona, Spain, 2011

[YD15] Ye, Q. ; Doermann, D.: Text Detection and Recognition in Imagery: A Survey. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37 (2015), July, Nr. 7, S. 1480–1500. http://dx.doi.org/10.1109/TPAMI.2014.2366765. – DOI 10.1109/TPAMI.2014.2366765. – ISSN 0162–8828

[Zhu05] Zhu, Xiaojin: Semi-Supervised Learning Literature Survey. (2005), Nr. 1530. http://pages.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf