Segmentierung und Strukturbasierte Adaptive Erkennung von Gebrauchsschrift in Historischen Dokumenten

Dissertation

zur Erlangung des akademischen Grades

Doktoringenieur (Dr.-Ing.)

angenommen durch die Fakultät für Informatik
der Otto-von-Guericke Universität Magdeburg

von: Dipl.-Inf. Markus Feldbach
geboren am: 27.11.1974 in Magdeburg

Gutachter:

Prof. Dr. Klaus-Dietz Tönnies
Prof. Dr. Jürgen Saedler
Prof. Dr. Xiaoyi Jiang

Promotionskolloquium: Magdeburg, den 09.02.2006

Zusammenfassung

Sollen Informationen aus historischen Aufzeichnungen mit Hilfe von Computern gewonnen werden, ist ein geeignetes Verfahren zur Schrifterkennung notwendig. Die Besonderheiten alter Dokumente ergeben sich aus den Umständen ihrer Entstehung. So ist das Papier häufig vergilbt und beispielsweise durch Stockflecke verunreinigt. Eine enge Schreibweise führt zu Störungen benachbarter Worte und Zeilen. Das Trainieren eines Erkenners ist schwierig, da dafür ein größerer Datensatz erforderlich ist, der aus den vorliegenden Dokumenten nicht oder nur sehr schwer gewonnen werden kann. Die Anpassung des Systems auf einen neuen Schreiber muss ohne Training erfolgen.

Es wird ein System vorgestellt, das auf der Basis digitalisierter Seiten von Kirchenbüchern die Zeilen segmentiert, Hypothesen über die Grenzen von Ziffern und Worten eines ausgewählten Bereiches erstellt und diese Objekte erkennt. Da hierbei ein struktureller Ansatz zur Anwendung kommt, ist ein Training nicht erforderlich. Eine Anpassung auf eine Schrift kann automatisch oder manuell erfolgen. Die Robustheit des Verfahrens sowie die Möglichkeiten der Anpassung wurden anhand der Datumsangaben in Kirchenbüchern des 18. und 19. Jahrhunderts getestet.

Danksagung

An dieser Stelle möchte ich mich bei all denjenigen bedanken, die mich bei der Erstellung dieser Arbeit unterstützten bzw. dies überhaupt ermöglichten. Insbesondere gilt dies für die Mitarbeiter des Instituts für Simulation und Graphik, die mehr als die technischen Voraussetzungen für das Entstehen dieser Arbeit schafften.

Für die hervorragende Betreuung während der gesamten Zeit danke ich Herrn Prof. Dr. Klaus-D. Tönnies. Seien es nun Fragen zur wissenschaftlichen Argumentation oder das kurzfristige Korrekturlesen eines Papers kurz vor dem Abgabetermin – er stand stets mit Rat und Tat zur Seite.

Ein herzliches Dankeschön geht an Herrn Dr. Helmut Mewes. Mit seinem Interesse auf dem Gebiet der Ahnenforschung und seinem Bestreben, den Computer für die Genealogie nutzbar zu machen, gab er den Anstoß für diese Arbeit. Durch ihn und die Kirchengemeinde Wegenstedt mit Pfarrerin Irene Heinecke wurde es mir ermöglicht, die Kirchenbücher der Gemeinde zu digitalisieren und somit Zugriff auf historische Schriften zu erhalten. Dafür vielen Dank. Gleiches gilt für die St. Johannis Gemeinde in Schönebeck / Bad Salzelmen mit ihrem damaligen Pfarrer Günter Schlegel. Auch hier konnte ich Seiten der Kirchenbücher digitalisieren.

Weiterhin möchte ich mich bei der Firma Graphikon mit ihrem Geschäftsführer Prof. Dr. Jürgen Saedler bedanken. Zur Implementierung wurde mir der Quellcode ihrer kommerziellen Software überlassen.

Auf jeden Fall danke ich Dr. Melanie Aurnhammer, die in atemberaubender Geschwindigkeit ihre Promotion abschloss, und Karsten Rink, der sowohl inhaltlich als auch formell die hier vorliegenden Seiten von größeren und kleineren „Unebenheiten“ befreite. Wenn es um die Erstellung englischsprachiger Veröffentlichungen ging, waren beide oft die Retter in letzter Not.

Schließlich geht ein großer Dank an meine Eltern. Sie unterstützten mich in jeder Hinsicht in dem recht langen Zeitraum der Erstellung dieser Arbeit und trugen so maßgeblich zu deren Fertigstellung bei.

Wichtige Formelzeichen

Abschnitt Zeilenfindung/-segmentierung

h_{Min}	Höhe der Minuskel
w_{Zei}	Mittlere Breite der Buchstaben
$d^{\mathrm{v}}_{\mathrm{y,max}}$	Maximaler vertikaler Versatz der lokalen Minima
$d^{\mathrm{v}}_{\mathrm{x,max}}$	Maximaler horizontaler Versatz der lokalen Minima
$d^{\mathrm{s}}_{\mathrm{y,max}}$	Maximaler vertikaler Versatz der Basisliniensegmente
$d^{\mathrm{s}}_{\mathrm{x,max}}$	Maximaler horizontaler Versatz der Basisliniensegmente

Abschnitt Wortsegmentierung

w_{date}	Breite des kompletten Datums
N_{pb}	Anzahl potentieller Wortgrenzen

Abschnitt Ziffernerkennung

S	Startpunkt eines Kreisbogens
E	Endpunkt eines Kreisbogens
H	Halbpunkt eines Kreisbogens
M	Mittelpunkt eines Kreisbogens
r	Radius eines Kreisbogens
l	Länge des Kreisbogens
σ	Richtung der Strecke $\overline{SM}$
ϵ	Richtung der Strecke $\overline{EM}$
α	Winkel des Kreisbogens
P	Prototyp
C	Kandidat
$\mathrm{size}\,(Prim)$	Größe des Primitivs $Prim$
$\mathrm{w}\,(Prim)$	Breitenfunktion, horizontale Ausdehnung von $Prim$
$\mathrm{h}\,(Prim)$	Höhenfunktion, vertikale Ausdehnung von $Prim$
$Prim_x$	Primitiv mit Index x
$Prim^{\mathrm{P}}_i$	Primitiv mit Index i eines Prototypen
$Prim^{\mathrm{C}}_j$	Primitiv mit Index j eines Kandidaten
$Prim_{\mathrm{FA}}$	Primitive der ersten Approximation
s	Länge des Skelettsegments
px_x	Skelettpixel mit Index x

BL	Basislinie
ML	Mittellinie
$\overline{dist_{\mathrm{PB}}}$	Mittlerer Abstand zwischen den potentiellen Wortgrenzen
th_{P}	Maximal zulässiger Abstand zwischen Primitiven für eine Fusion
$d_{\max}$	Maximaler Approximationsfehler zwischen Skelett und Primitiv
m	Korrespondenz (Matching) zwischen Prototyp und Kandidat

Abschnitt Ziffernerkennung, Kostenberechnung

c^{cipher}	Gesamtkosten einer Hypothese der Ziffernerkennung
c^{ap}	Approximationskosten zwischen Primitiv und Stichverlauf
c^{m}	Matchingkosten zwischen Prototyp und Kandidat
c^{us}	Kosten nicht genutzter Striche der Ziffernerkennung
c^{tr}	Translationskosten
c^{sc}	Skalierungskosten
c^{ro}	Rotationskosten
c^{sh}	Krümmungskosten
c^{wp}	Kreisteilkosten
nv	Normierungswert
f^{ap}	Faktor der Approximationskosten
f^{tr}	Faktor der Translationskosten
f^{sc}	Faktor der Skalierungskosten
f^{ro}	Faktor der Rotationskosten
f^{sh}	Faktor der Kosten der Formabweichung
f^{wp}	Faktor der Kosten für falschen Kreisteile
$f_{\mathrm{z}}^{\mathrm{us}}$	Faktor der Kosten nicht genutzter Striche der Ziffern

Abschnitt Worterkennung

$f_{\mathrm{w}}^{\mathrm{us}}$	Faktor der Kosten nicht genutzter Striche der Wörter
c^{word}	Gesamtkosten einer Hypothese der Worterkennung
c_x^{sect}	Kosten des Abschnitts Nr. x
f^{horiz}	Faktor der Kosten der horizontalen Translation eines Abschnitts
f^{vert}	Faktor der Kosten der vertikalen Translation eines Abschnitts

Glossar

Basislinie	Textlinie, bestimmt durch die untere Ausdehnung der Buchstaben (auch Grundlinie)
Fusionieren	Erzeugen von ein oder zwei Primitiven aus zwei kleineren; siehe Seite 48 ff.
Gebrauchsschrift	Flüchtigere Schrift der vergangenen Jahrhunderte im Gegensatz zur strengen Buch- und Urkundenschrift
Genealogie	Familienforschung, Historische Hilfswissenschaft
Hauptprimitiv	Ausgewähltes Primitiv, das zur Bestimmung von Translation und Skalierung genutzt wird; siehe Seite 51
Majuskel	Großbuchstabe
Minuskel	Kleinbuchstabe
Mittellinie	Textlinie, die von der oberen Ausdehnung der kleinen Buchstaben bestimmt wird
Korrespondenz	siehe Seite 51 f.
Ligatur	Verbindung zwischen zwei Buchstaben
Oberlänge	Striche der Schrift, die die Mittellinie nach oben überschreiten
Paläografie	Historische Hilfswissenschaft zur Bestimmung der räumlichen und zeitlichen Verwendung von Schriften
Primitiv	Kreisbogen, siehe Seite 45 ff.
Schaft	Senkrechter Grund- oder Hauptstrich eines Buchstabens
Schriftform	siehe Seite 10 f.
Schriftklasse	siehe Seite 10 f.
Schriftstil	siehe Seite 10 f.
Schulterstrich	Querstrich eines Buchstabens in Höhe der Mittellinie, z. B. beim kleinen „t“
Unterlänge	Striche der Schrift, die die Basislinie nach unten überschreiten
Zeilenhauptraum	Bereich einer Zeile zwischen Basis- und Mittellinie

Inhaltsverzeichnis

Kapitel 1

Einleitung

Die Schrift hat das Geheimnisvolle, dass sie redet.

Paul Claudel (1868-1955)

Die Fähigkeit des Menschen, Schrift lesen zu können, ist heutzutage so selbstverständlich, dass diese enorme kognitive Leistung kaum Beachtung findet. In Bruchteilen einer Sekunden können die Schriftobjekte, die sich meist durch Helligkeit oder Farbe vom Hintergrund abheben, in Zeilen, Buchstaben und Worte eingeteilt und einer Bedeutung zugeordnet werden. Nur jene Personen, die sich mit der automatischen Erkennung von Schrift beschäftigen, können die Komplexität dieser kognitiven Fähigkeit richtig einschätzen. Umso höher ist die Tatsache zu bewerten, *„daß trotz der hohen Variabilität zwischen verschiedenen Handschriften es dennoch möglich ist, handschriftliche Notizen anderer Menschen relativ einfach zu lesen“* [103].

Die automatische Erkennung von Druckschrift kann mittlerweile als ausgereift betrachtet werden. Die günstigen Eigenschaften wie getrennte Buchstaben in konstanter Form und Größe, die sich auf geraden und voneinander getrennten Zeilen befinden, sind hier von Vorteil. Solange auftretende Störungen ein bestimmtes Maß nicht übersteigen, arbeiten die Erkennungssysteme nahezu fehlerfrei.

Sind handgeschriebene Texte der Gegenstand einer automatischen Erkennung, so sind heutige Verfahren noch nicht ohne Weiteres in der Lage, verlässliche Resultate zu liefern. Zu groß sind die Unterschiede, wie Menschen schreiben – Mehrdeutigkeiten entstehen. Lässt sich der Kontext jedoch eingrenzen, können entsprechende Erkennungssysteme auch handschriftliche Texte lesen. Beispielsweise werden bei vielen Paketunternehmen und Versandbetrieben die Adressen auf Briefen und Paketen zu einem Großteil automatisch erkannt. Dies ist möglich, da bekannt ist, welche Information an einer bestimmten Stelle zu finden ist. Für kleinere Lexika lag die Erkennungsrate bereits vor 10 Jahren bei 90 bis 95 % (100 bzw. 10 Worte) [43]. Allerdings basieren all diese Verfahren auf einer statistischen Auswertung vieler tausender Testdaten. Es besteht keine Möglichkeit, Wissen über Charakteristika einer Schrift zu nutzen oder Fehlleistungen während der Erkennung detailiert zu begründen.

1.1 Motivation

Die Gründe, warum ein Schriftstück erstellt wird, sind verschieden. In einem Fall ist es das Ziel, bestimmte Informationen für sich selber oder für andere Personen aufzubewahren. In einem anderen Fall sollen Informationen über eine große Distanz transportiert werden. Diese Arbeit befasst sich mit alten Kirchenbüchern als Dokumente einer Klasse, die zur Konservierung von Informationen geschaffen wurden. Es ist das Ziel des Schreibers, sich selber oder anderen Personen in der Zukunft die Möglichkeit zu geben, durch das Lesen dieser Texte diese Informationen wieder zu gewinnen.

Personen, die heute an solchen alten Texten interessiert sind, sind Historiker oder Genealogen. Sie gewinnen aus verschiedenen Quellen Informationen, um die Vergangenheit unserer Gesellschaft zu rekonstruieren. Neben Dokumenten aus Kanzleien der Gerichte, Städte und Steuerämter wie z. B. Gerichts- und Grundbücher, Notariatsakten, Kopialbücher[1] und Ratsprotokolle sind auch oder gerade Kirchenbücher vor allem im ländlichen Bereich eine sehr ergiebige Quelle. Während heute wichtige Daten immer seltener in Papierform und viel häufiger elektronisch erfasst und verwaltet werden, gab es diese Möglichkeit in den vergangenen Jahrhunderten natürlich nicht. Die kirchliche Gemeinde stand im Mittelpunkt gesellschaftlicher Ereignisse. Über Jahrhunderte hinweg wurden so die wichtigsten Geschehnisse der Menschen wie Geburt, Heirat und Tod festgehalten. Durch zeitaufwändiges Recherchieren in Archiven können so Zusammenhänge zwischen Personen der Vergangenheit hergestellt und ganze Stammbäume rekonstruiert werden. Dies ist das Anliegen der Genealogen. Sehr viele Menschen sind heutzutage auf der Suche nach ihren Wurzeln. Sie möchten wissen, wer ihre Vorfahren waren, wo sie lebten und dies für einen möglichst großen Zeitraum. Vereine wurden gegründet und spezielle Computerprogramme zur Verwaltung von Verwandschaftsverhältnissen entstanden [92]. Es ist jedoch sehr aufwendig, Informationen aus diesen alten Dokumenten zu gewinnen. Die alte Schrift ist schwer zu lesen und Menschen, die dies problemlos können, sind rar. Daher wäre es wünschenswert, wenn von Seiten der Informatik ein Werkzeug entstünde, damit auch weniger erfahrene Personen den Inhalt alter Kirchenbücher erschließen können. Es ist abzusehen, dass es keine vollautomatische Lösung geben wird. Einerseits sind der Kreativität mancher Pfarrer bzw. Schreiber einer Gemeinde keine Grenzen gesetzt, andererseits kommt es hin und wieder zu Irrtümern und Schreibfehlern, die nur durch einen Menschen erkannt werden können. Denn nur ein Mensch kann flexibel genug auf unvorhersehbare Situationen reagieren, ein Computer wird dazu in absehbarer Zeit nicht in der Lage sein.

Dennoch ist es sinnvoll, zu untersuchen, inwieweit eine Unterstützung durch den Rechner bei der Transkription alter Dokumente möglich ist. Es existiert sowohl Wissen über das Aussehen der Schrift als auch über die Struktur alter Dokumente. Die Paläografie als historische Hilfswissenschaft befasst sich mit den Merkmalen der Handschrift im zeitlichen und regionalen Kontext. Seit einiger Zeit werden in diesem Bereich Untersuchungen durchgeführt, um die zeitliche Entwicklung handge-

[1] Bücher mit Abschriften (Kopien) von Urkunden und Verträgen.

schriebener Buchstaben mit Hilfe von Software und struktureller Analyse zu beschreiben [15]. Es wird dabei zwischen Struktur und Form unterschieden. Die Struktur umfasst die räumliche Anordnung der Striche und die Proportionen des Buchstabens während die Form z. B. dekorative Elemente wie Serifen beinhaltet.

Die Unterschiede, die zwischen der Schrift zweier Schreiber besteht, oder der Schrift eines Schreibers im Laufe der Zeit, werden so systematisiert und sind teilweise durch Parameter zu beschreiben. Diese Informationen können genutzt werden, um das Erkennungssystem auf eine neue Schrift einzustellen, ohne ein Training durchführen zu müssen. Paläografisches Wissen zur automatischen Erkennung alter Schriften kann auf diese Weise genutzt werden.

1.2 Gegenstand der Arbeit

Es stellt sich die Frage, inwieweit die Schrift alter Dokumente automatisch verarbeitet werden kann. Es sind die verwendeten Materialien, die Schreibtechnik und die Art der Schrift, die die Unterschiede zu modernen handschriftlichen Aufzeichnungen ausmachen.

Die besonderen Merkmale alter Dokumenten sind der Grund für die Schwierigkeiten, die einer automatischen Texterkennung im Wege stehen. Das häufige Auftreten von Artefakten, die großen Varianzen der Schrift sowie das Fehlen eines ausreichend großen Trainingsdatensatzes verhindern die Verwendung bestehender Verfahren zur Handschriftenerkennung. Die Zahl der Kirchenbücher in einer Gemeinde ist nicht groß genug, um genügend Trainingsdaten für eine statistische Auswertung zu erzeugen.

Das hier vorgestellte Erkennungssystem soll daher die erforderlichen Eigenschaften erfüllen:

- Erkennung von handgeschriebenen Wörtern ohne umfassendes Training.
- Manuelle und automatische Anpassung des Systems an eine neue Schrift.
- Robustheit gegenüber Artefakten, die durch das Material und den Schriftstil verursacht wurden.
- Transparentes Verhalten des Erkenners im Vergleich zu statistisch trainierten Systemen.

Es wäre wünschenswert, den Text der Kirchenbücher Seite für Seite automatisch zu erkennen. Doch dieses Ziel ist bei dem momentanen Stand der Technik unrealistisch. Das Problem muss eingeschränkt werden. Daher ist das Datum in den Kirchenbucheinträgen der Gegenstand dieser Arbeit. Zum Einen gibt es nur eine begrenzte Zahl von Variationen, wie sich solch eine Angabe zusammensetzen kann. Zum Anderen stellt das Datum einen Teil der besonders wichtigen Informationen dar, die für Historiker und Genealogen interessant sind. Auch unter diesen eingeschränkten Bedingungen sind Aussagen bezüglich der oben genannten Eigenschaften möglich.

1.3 Aufbau der Arbeit

In dem folgenden **2. Kapitel** werden die Merkmale alter Kirchenbücher beschrieben, die das besondere Vorgehen zur Vorverarbeitung, Segmentierung und Erkennung notwendig machen. Neben den Materialien alter Dokumente sind es vor allem die Schriften, die eine angepasste Verarbeitung erfordern. Die Erkenntnisse der Paläografie, welche sich mit dem Aussehen und der Entwicklung der Schrift in der Vergangenheit beschäftigt, kann genutzt werden, um eine Erkennung ohne Training durchzuführen.

In **Kapitel 3** wird der Aufbau eines Erkennungssystems erläutert. Ein Dokument muss digitalisiert werden bevor eine informationstechnische Verarbeitung erfolgen kann. Durch eine Vorverarbeitung werden die gewonnenen Daten für die Segmentierung und Erkennung vorbereitet. Segmentierung heißt Aufteilung der Schriftobjekte in Zeilen und Aufteilung in Ziffern und Wörter. Nach diesem Schritt kann die Erkennung eines Objektes vorgenommen werden. Es wird untersucht, welche Ansätze für die einzelnen Module existieren und welche Möglichkeiten sie zur Lösung des vorliegenden Problems bieten.

Das Verfahren zur Wortsegmentierung am Beispiel des Datums wird in **Kapitel 4** beschrieben. Die Bewertung der Lücken zwischen Textobjekten ist nicht ausreichend, da es beispielsweise zu Berührungen benachbarter Wörter kommen kann. A-priori-Wissen wird eingebracht, um Hypothesen über die Positionen potentieller Wortgrenzen zu erzeugen.

Der strukturelle Ansatz zur Erkennung der Ziffern wird in **Kapitel 5** behandelt. Dieses an Kahn et al. [37, 38] angelehnte Verfahren basiert auf der Generierung einer strukturellen Repräsentation der Striche. Anhand von Prototypen werden Ähnlichkeiten zu den Erscheinungsformen von Ziffern bestimmt.

In **Kapitel 6** wird das strukturelle Erkennungsverfahren erweitert, um Wörter holistisch zu erkennen. Die Umsetzung erfolgt anhand der Erkennung der Monatsnamen. Das Verfahren bietet die Möglichkeit, ohne Training eine Anpassung an eine neue Schrift durchzuführen. Dies kann manuell oder automatisch erfolgen.

Das **Kapitel 7** dient zur Darstellung der Ergebnisse, die durch Tests und Experimente mit den zur Verfügung stehenden Daten entstanden. Sie demonstrieren die erfolgreiche Anwendung der in den vorhergehenden Kapiteln beschriebenen Techniken.

Abschließend wird in **Kapitel 8** eine Zusammenfassung der Arbeit gegeben und beleuchtet, inwieweit das hier beschriebene System erweitert und verbessert werden kann.

Kapitel 2

Kirchenbücher als Informationsquelle

Bücher sind der geschätzte Reichtum der Welt,
die richtige Erbschaft von Generationen und Völkern.

Henry David Thoreau (1817-1862)

Seit dem Beginn des 17. Jahrhunderts wurden wichtige Ereignisse in einer Kirchengemeinde vom Pfarrer in so genannten Kirchenbüchern oder Matrikeln festgehalten. In der Regel beginnen die Aufzeichnungen allerdings erst 1648, da während des Dreißigjährigen Krieges keine Eintragungen getätigt oder die schon existierenden Bücher zerstört wurden. Je nach Art des Ereignisses erfolgten die Eintragungen in Tauf-, Heirats- oder Totenbücher. Der Umfang eines solchen Eintrags schwankt. Es kann der Wohnort und Beruf der beteiligten Personen angegeben sein, zumindest enthält er das Datum und den oder die Namen der Personen.

Die für diese Arbeit gewonnenen Daten sind digitalisierte Seiten von Kirchenbüchern der Gemeinde Wegenstedt. Um über eine gewisse Auswahl zu verfügen, wurden Dokumente um 1720, 1770 und 1810 digitalisiert. Desweiteren wurden Kirchenbücher der St. Johannis Gemeinde in Schönebeck-Bad Salzelmen um 1770 digitalisiert, um das Spektrum der Schriften weiter zu erhöhen.

2.1 Material

Für die hier vorliegenden Dokumente wurde Papier als Beschreibstoff genutzt, das vom 14. bis zum 17. Jahrhundert das teurere Pergament vollständig in Europa verdrängt hatte. Die Art der Lagerung und die Nutzung der Bücher hinterließ über die Jahre Spuren, die die Erkennung der Schrift erschweren. Eine kühle und feuchte Lagerung führt häufig zu Schimmelbildung, die durch Stockflecken sichtbar wird (siehe Abbildung 2.1). Einen durch Eisengallustinte verursachten Tintenfraß, der bei Dokumenten dieser Zeit auftreten kann, konnte hingegen nicht beobachtet werden.

Entsprechende Filter in der Vorverarbeitung reduzieren die niederfrequenten Artefakte. Dennoch muss das Erkennungsverfahren robust gegenüber Verunreinigungen solcher Art sein.

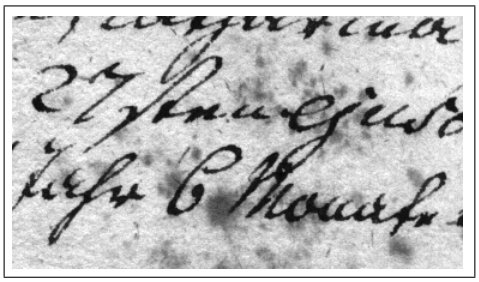


Abbildung 2.1: Artefakte wie Stockflecken erschweren den Erkennungsprozess.

2.2 Die Seite eines Kirchenbuchs

Fand in einer Gemeinde eine Taufe, eine Heirat oder eine Beerdigung statt, so wurde dieses Ereignis in der Regel zeitnah in dem entsprechenden Buch festgehalten. Daher ist davon auszugehen, dass die Eintragungen in einer chronologischen Reihenfolge vorliegen. In seltenen Fällen können auch nachträgliche Ergänzungen oder Korrekturen beobachtet werden. Diese stellen für eine automatische Verarbeitung eine zusätzliche Erschwernis dar, sollen hier aber nicht weiter betrachtet werden.

Mit der Zeit entstanden so komplett beschriebene Seiten des Kirchenbuches, auf denen mehrere Einträge zu finden sind. Auch über Jahrhunderte hinweg hat sich dieses Erscheinungsbild nicht wesentlich geändert. Erst zum Ende des 19. Jahrhunderts wurde die tabellarische Form angewandt. Dieses Format soll hier jedoch nicht im Vordergrund stehen. Es kann davon ausgegangen werden, dass sich für eine Erkennung dieser Tabellen geringere Schwierigkeiten ergeben, da sich bei dieser Form die Positionen der einzelnen Informationen klarer bestimmen lassen [69].

2.3 Die Datumsangabe

Es gibt unterschiedliche Formen, wie Einträge in Kirchenbüchern getätigt wurden. Dies gilt auch für die Angabe des Datums. In einigen Büchern wurde das Datum nach dem römischen Kalender angegeben – inklusive der römischen Ziffern. Meistens dominierten jedoch Angaben in julianischer bzw. gregorianischer Kalenderrechnung und mit ihnen die Angabe der Tage mittels arabischer Ziffern. Sie bilden den Gegenstand des entwickelten Erkennungsverfahrens.

Unterschiede gibt es auch in der Art und Position des angegebenen Datums im Eintrag oder auf der entsprechenden Seite, d. h. wo sich die Bestandteile des Datums (Jahr, Monat, Tag) befinden.

In größeren Gemeinden in Städten ist die Zahl der Ereignisse relativ groß. Pro Monat kam es zu mehreren Geburten oder Todesfällen[1]. Daher wurde neben dem Jahr auch der Monat als Überschrift angeführt, sodass lediglich der Tag des Datums für einen Eintrag angegeben wurde.

[1] Vermählungen ereigneten sich um ein Vielfaches seltener als Geburten oder Todesfälle.

Der Hauptteil, der hier zur Verfügung stehenden Dokumente, stammt aus dem Dorf Wegenstedt, einer relativ kleinen Gemeinde. Die Jahreszahl wurde hier als Überschrift in dem Kirchenbuch eingetragen und die entsprechenden Ereignisse dieses Jahres darunter aufgeführt, jeweils mit Tag und Monat (siehe Abbildung 2.2).

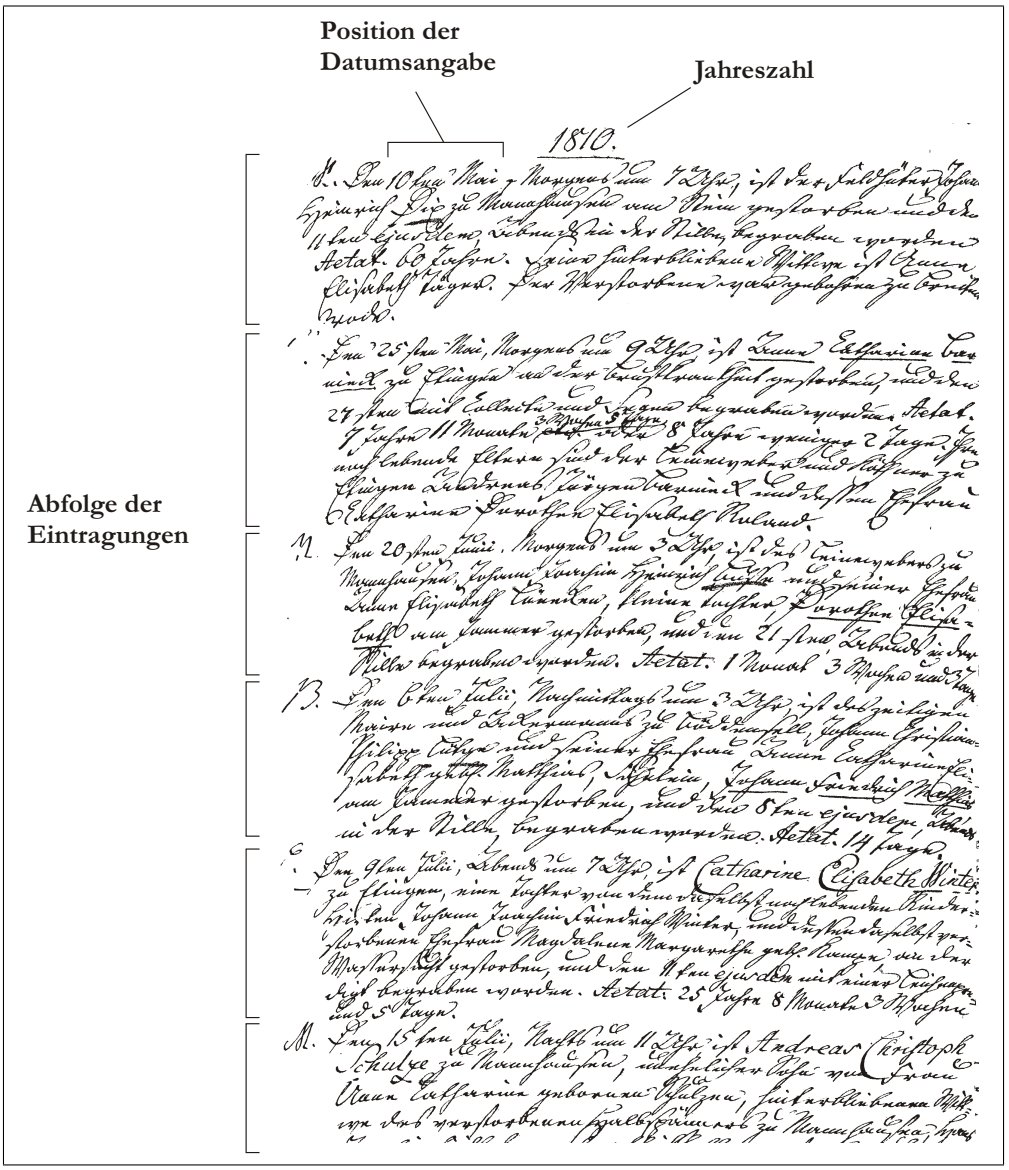
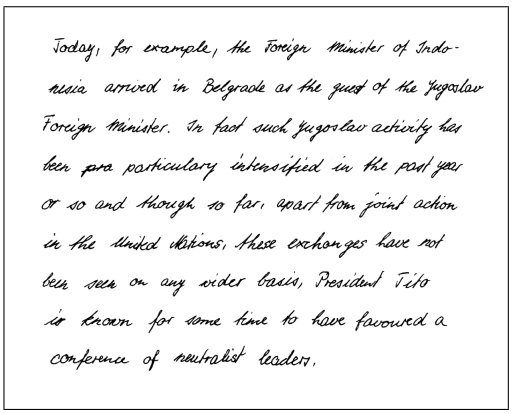


Abbildung 2.2: Seite eines Wegenstedter Totenbuchs aus dem Jahre 1810. Die Eintragungen wurden zeitnah und somit chronologisch getätigt. Das Datum befindet sich hier am Anfang eines Eintrags.

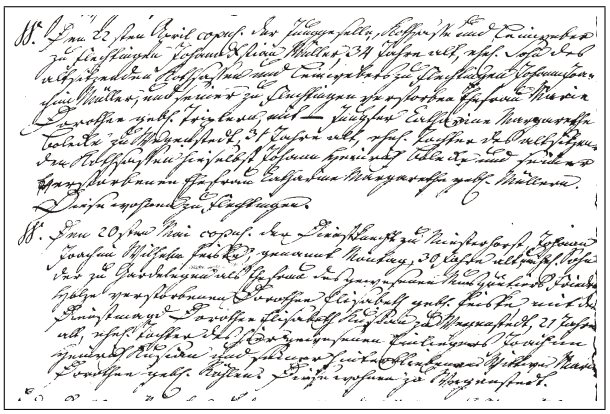
2.4 Besonderheiten der Schrift

Die Schrift der Kirchenbücher des 18. und 19. Jahrhunderts besitzt Besonderheiten, die berücksichtigt werden müssen, wenn eine Erkennung erfolgreich sein soll. Bestimmt werden diese Besonderhei-

ten zum Einen von der Schrift, die in der damaligen Zeit verwendet wurde, und zum Anderen von der Form des Dokuments. Einige Eintragungen wurden sehr schnell und unsauber vorgenommen. Im Vergleich dazu liegt bei vielen Arbeiten über Handschrifterkennung von allgemeinen Texten eine Schrift vor, die durch einen kooperativen Schreiber erzeugt wurde [57, 58] (siehe Abbildung 2.3(a)). Im Besonderen gilt dies für die Online-Erkennung. Bei Kirchenbüchern oder generell bei historischen Schriften muss hingegen von einem unkooperativen Schreiber gesprochen werden.


Today, for example, the foreign minister of Indonesia arrived in Belgrade as the guest of the Yugoslav Foreign Minister. In fact such Yugoslav activity has been par particularly intensified in the past year or so and though so far, apart from joint action in the United Nations, these exchanges have not been seen on any wider basis, President Tito is known for some time to have favoured a conference of neutralist leaders.

(a) Beispieltext aus Marti et al. [58].



(b) Ausschnitt einer Kirchenbuchseite (Wegenstedt, 1813).

Abbildung 2.3: Vergleich zwischen normaler und Kirchenbuchschrift. (a) Bisherige Erkennungssysteme setzen gut getrennte Zeilen und Wörter voraus. (b) Die enge Schreibweise ist eines der besonderen Merkmale alter Kirchenbücher.

2.4.1 Verlauf der Zeilen

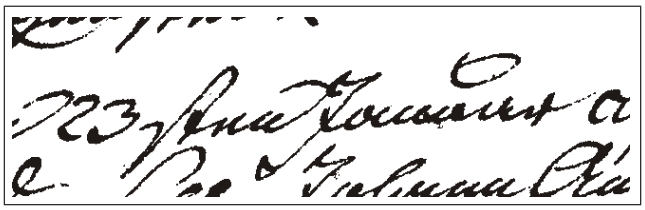
Textlinien sind die maximal vier Linien, die durch die vertikale Ausdehnung der Minuskel und Majuskel gebildet werden. Sowohl für die Segmentierung der Zeilen als auch für die Worterkennung muss berücksichtigt werden, dass diese Textlinien in den Kirchenbüchern nicht so stark ausgebildet sind, wie dies bei anderen, sorgfältiger geschriebenen Schriften wie Briefen der Fall ist. Ganz besonders die beiden äußeren Textlinien der Ober- und Unterlängen sind oft nur sehr schwach oder gar nicht ausgebildet. Es wird daher für die Zeilensegmentierung nur der Verlauf der beiden inneren Textlinien rekonstruiert.

Es wurde frei Hand auf die leeren Seiten des Kirchenbuches geschrieben. Hilfslinien gab es nicht. Daher ist der Verlauf der Zeilen weder parallel zu einander noch gerade. Es kann eine Bogenform vorliegen oder jedes Wort besitzt seine eigene verikale Position und Richtung.

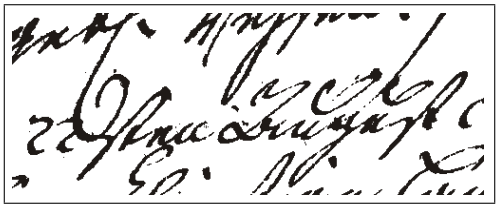
2.4.2 Wortberührungen

Wie auch in anderen handschriftlichen Aufzeichnungen sind in historischen Kirchenbuchaufzeichnungen die Grenzen zwischen zwei Worten meist durch eine Lücke in horizontaler Richtung zwi-

schen den Schriftobjekten gekennzeichnet. Darüber hinaus können aber auch Berührungen zwischen benachbarten Wörtern bestehen. Dies resultiert aus der engen Schreibweise, den schwungvoll geschriebenen Zeichen am Anfang und Ende sowie aus Strichen, die von benachbarten Zeilen stammen und zwei Worte in der aktuellen Zeile berühren (siehe Abbildung 2.4).



(a)



(b)

Abbildung 2.4: Berührung benachbarter Worte durch lang gezogenen Schlussstrich (a) oder Unterlängen darüber liegender Zeilen (b).

Unterlängen wurden oft sehr ausschweifend und schwungvoll geschrieben. Dadurch ragen sie in den Raum der nächsten Zeile. Im Bezug auf die Oberlängen besteht dieses Problem nicht. Vermutlich liegt dies an der Präsenz der vorherigen Zeile während des Schreibens, sodass dort nicht hinein geschrieben wurde. Schon die Existenz dieser Striche in einem anderen Zeilenraum erschwert die Segmentierung und Erkennung. Ein größeres Problem entsteht aber durch die Berührung von Textobjekten benachbarter Zeilen.

2.4.3 Art und Form der Schrift

Handschriftliche Aufzeichnungen neuerer Zeit unterliegen einer Reihe von Variationen, die im Wesentlichen auf individuelle Angewohnheiten zurückzuführen sind. Seit Mitte des 20. Jahrhunderts wird in Deutschlands Grundschulen die lateinische Schreibschrift gelehrt. Im Vergleich dazu hat sich die Schreibschrift der vergangenen Jahrhunderte stetig verändert. Die Paläografie ist die Wissenschaft, die sich mit diesen Veränderungen und zeitlich-regionalen Zusammenhängen beschäftigt, um eine systematische Betrachtung der Merkmale einer Schreibschrift zu ermöglichen.

Basierend auf der fünfstufigen Kategorisierung von Schriftprodukten von Petrucci [74] lassen sich drei wesentliche Kategorien ableiten, die im Folgenden mit eindeutigen Begriffen bezeichnet werden:

Schriftklasse: Das verwendete Alphabet legt das grundlegende Aussehen eines Buchstabens und damit eines Wortes fest. Beispielsweise Lateinisch, Kyrillisch, Gotisch.

Schriftform: Für ein Alphabet gibt es unterschiedliche Ausprägungen. Beispielsweise Gotische Kursive, Humanistische Kursive, Lateinische Kurrentschrift.

Schriftstil: Form und Gestalt der Buchstaben unterliegen individuellen Ausprägungen und werden u. a. durch das verwendete Schreibgerät beeinflusst.

Diese Begriffsunterscheidung ist eng verbunden mit den Faktoren, die das Aussehen einer Handschrift bestimmen. Neben dem verwendeten Schreibgerät sind dies:

Zeitliche und regionale Gegebenheiten. Für jede Region (Großbritannien, Deutschland, Italien, Nordamerika,...) gibt es für jede Zeit bestimmte Schriftklassen und Schriftstile.

Form des Dokuments. In Abhängigkeit des jeweiligen Dokuments variiert die Schrift in ihrer Qualität. So spricht man einerseits von der strengen Buch- oder Urkundenschrift und andererseits von der flüchtigen Gebrauchsschrift wie z. B. in Kirchenbüchern [26].

Individuelle Ausprägungen. Die Handschrift eines Menschen besitzt individuelle Merkmale, mit denen sogar eine eindeutige Identifizierung des Schreibers möglich ist [62, 59].

Im Vergleich zur heutigen Schreibschrift unterliegen die Schriften des 18. und 19. Jahrhunderts höheren Variabilitäten, verursacht durch die Verwendung unterschiedlicher Schreibformen. Bis zum Beginn des 18. Jahrhunderts waren es die Schreibmeister der verschiedenen Schreib- und Meisterschulen, die die Ausprägungen der verwendeten Kurrentschriften beeinflussten [99]. Danach wurde das Schreiben als Grundlehrfach in den allgemeinen Schulen eingeführt. Von 1714 an wurde in Preußen die Schrift des Schreibmeisters Hilmar Curas gelehrt, die fast alle deutschsprachigen Länder beeinflusste. Im Jahre 1817 wurde sie durch die Berlinischen Schulvorschriften von Hennig modernisiert.

Wendet man diese paläografischen Erkenntnisse auf die automatische Schrifterkennung alter Texte an, kann für die Auswahl der Merkmale folgendes festgestellt werden: Grundsätzlich stellen sämtliche Striche die Merkmale eines Wortes dar. Doch besitzt nicht jeder Strich die gleiche Bedeutung. Für ein Wort in einer bestimmten *Schriftform* existieren eine Reihe von wesentlichen Merkmalen, die sich in Form von bestimmten Strichen oder Strichkombinationen darstellen. Sowohl die Veränderungen der Schrift eines Schreibers über die Zeit als auch die Abweichungen zwischen unterschiedlichen Schreibern, die aber die gleiche *Schriftform* benutzen, beeinflussen nicht die Existenz dieser wesentlichen Merkmale. Diese Variationen betreffen lediglich die Ausprägung dieser Striche. Parameter wie Größe, Position und Rotation können in einem bestimmten Rahmen variieren. Beim Vergleich unterschiedlicher *Schriftklassen* hingegen stellt man schnell fest, dass hier vollkommen unterschiedliche Merkmale vorliegen – einzelne Buchstaben vollkommen anders gezeichnet werden.

Ziffern in alten Dokumenten

Im Vergleich zu den deutschen oder lateinischen *Wörtern* hat sich das Aussehen der arabischen *Ziffern* im Laufe der Zeit kaum geändert. Lediglich vor 1500 gab es Unterschiede vor allem bei den Ziffern 4, 5 und 7 [82, 26]. Für den hier betrachteten Zeitraum spielt dies jedoch keine Rolle. Es gibt jedoch Unterschiede der Bewegungsabläufe, die zu Variationen in der Form führen können (siehe Abbildung 2.5).

Abbildung 2.5: Unterschiedliche Bewegungsabläufe während des Schreibens führen zu Variationen der Schrift [62].

Kapitel 3

Überblick und Einordnung der Arbeit

Nichts gewinnt so sehr durch das Alter wie
Brennholz, Wein, Freundschaften und Bücher.

Francis Bacon (1561-1626)

Viele Schritte sind notwendig, um den Text eines Dokumentes informationstechnisch zu verarbeiten. In diesem Kapitel wird dargestellt, aus welchen Teilen ein System zur Schrifterkennung besteht und wie diese Teile zusammenarbeiten. Dabei wird herausgestellt, inwieweit es erforderlich ist, aufgrund der Charakteristiken alter Schrift im Bezug auf existierende Verfahren neue Ansätze zu verfolgen bzw. bestehende zu erweitern. Der Schwerpunkt dieser Arbeit liegt in der Segmentierung und Erkennung von Ziffern und Wörtern. Dieses Kapitel verdeutlicht, wie sich diese Module in das System eingliedern.

3.1 Überblick des Erkennungssystems

Eine Kirchenbuchseite enthält vielfältige Informationen: Farbe und Farbnuancen des Papiers und der Schrift, Art und Stärke von Verunreinigungen, Beschädigungen des Papiers aber auch Charakteristika der Schrift und vor allem der Inhalt des Geschriebenen. Für die Schrifterkennung steht die Information im Vordergrund, die bewusst vom Schreiber hinterlassen wurde. Es sind letztendlich die Form und Anordnung der Striche, die er mit dem Schreibgerät auf das Papier brachte. In mehreren Stufen wird die Datenmenge reduziert, ohne dabei diese Information zu verlieren.

Der grundlegende Ablauf eines Erkennungsprozesses ist durch die Abfolge **Digitalisieren, Vorverarbeiten, Segmentieren, Klassifizieren** gekennzeichnet. Dies gilt sowohl für die On-line- [32, 100] als auch für die Off-line-Erkennung [41, 34]. Bei der On-line-Erkennung erfolgt der erste Schritt durch das Digitalisieren der Stiftposition über die Zeit [75]. Durch das Erfassen der zeitlichen Abfolge ist eine Zuordnung der Striche zu Zeilen, Wörtern und Buchstaben sehr viel einfacher als bei der Off-line-Erkennung. Hier erfolgt das Digitalisieren durch die optische Abtastung einer Oberfläche –

in der Regel eines Blatts Papier. In den meisten Verfahren erfolgt anschließend eine Zuordnung der resultierenden Bildpunkte zu Vorder- und Hintergrund. Nur wenige Ansätze existieren, die auf der Basis der Grauwertbilder Merkmale zur Erkennung gewinnen [1, 98]. Dies soll hier aber nicht weiter betrachtet werden. Es ist gerade das Ziel einer strukturellen Verarbeitung, die Menge der Daten zu reduzieren und in jedem Schritt der Vorverarbeitung eine höherwertige semantische Repräsentation zu erzeugen.

Ein entscheidender Nachteil der Off-Line-Erkennung ist das Fehlen der zeitlichen Auflösung. Es existiert lediglich das zweidimensionale Bild der durch das Schreibgerät aufgetragenen Tinte. Zwar existieren Verfahren zur Rekonstruktion der Strichabfolge [48, 33], jedoch arbeiten diese Verfahren nicht fehlerfrei und benötigen A-priori-Wissen. Hinzu kommt, dass Artefakte durch Papier und Verunreiningungen diese Rekonstruktion erschweren. Daher werden diese Ansätze in dieser Arbeit nicht weiter betrachtet. Dennoch soll an dieser Stelle nicht ausgeschlossen werden, dass eine Rekonstruktion der Strichreihenfolge die Erkennung der Schrift alter Dokumente verbessern kann.

Zunächst ist eine angepasste Vorverarbeitung zur Einteilung der Bildpunkte in Vorder- und Hintergrund erforderlich. Anschließend müssen die daraus entstehenden Vordergrundobjekte zunächst einer Zeile und anschließend einem Wort zugeordnet werden. Aktuelle Verfahren zur allgemeinen (*unconstrained*) Schrifterkennung setzen oftmals Schrift voraus, die sich durch gut getrennte Zeilen und Wörter auszeichnen [58, 34]. Dies ist bei Kirchenbüchern nicht der Fall und erfordert eine entsprechende Verarbeitung zur Zeilen- und Wortsegmentierung.

Zur Zeilensegmentierung wird das in [17] vorgestellte Verfahren genutzt, das die meisten Schriftobjekte eindeutig einer Zeile zuordnet. Für die Wortsegmentierung ist dies nicht ohne Weiteres möglich. Die geometrischen Merkmale reichen für eine sichere Einteilung einer Zeile in ihre Wörter nicht aus. Für den hier betrachteten Fall der Erkennung des Datums werden daher mehrere Hypothesen über die Position von Wortgrenzen aufgestellt. Für die Hypothesen mit den höchsten Wahrscheinlichkeiten werden die potentiellen Ziffern und Wörter (in diesem Fall der Monatsname) durch die entsprechenden Erkennungsmodule erkannt. Für ein besseres Verständnis dieses Ablaufs befindet sich am Ende dieses Kapitels auf Seite 34 die Abbildung 3.12, die das Zusammenspiel von Segmentierung und Erkennung verdeutlicht.

Es existieren unterschiedliche Ansätze, auf welche Weise eine Klassifizierung erfolgen kann. Statistisch arbeitende Verfahren zeichnen sich oft durch eine Verarbeitung einer relativ hohen Zahl einfacher Merkmale aus [90]. Es sind Merkmale, die unabhängig von der Klasse für jeden Kandidaten extrahiert werden können, da ein statistischer Klassifikator einen Merkmalsvektor mit konstanter Länge voraussetzt. Die Klassifikatoren müssen durch einen großen Datensatz trainiert werden. Strukturell basierte Verfahren hingegen analysieren komplexere Merkmale wie den Verlauf der Striche [7, 38, 22]. Bei der Worterkennung wird zusätzlich zwischen segment-basiert und holistisch unterschieden. Die erste Variante zerlegt das Wort in kleinere Bestandteile. Aus der sich daraus ergebenden Abfolge von zuvor trainierten Merkmalen werden Hypothesen über das Wort gebildet. Bei

holistischen Verfahren erfolgt keine Aufteilung des Wortes. Analysiert werden relativ wenige aber markante komplexere Merkmale des Wortes wie Ober- und Unterlängen.

Das Digitalisieren, die Vorverarbeitung sowie die Segmentierung der Zeilen wurden ausführlich in [17], [18] und [19] beschrieben. Diese Bereiche werden in dieser Arbeit daher nur kurz erwähnt und durch aktuelle Informationen und Erkenntnisse ergänzt. Anschließend erfolgt eine Darstellung der Module der Wortsegmentierung sowie der Ziffern- und Worterkennung im Bezug auf existierende Verfahren und dem eigenen Ansatz.

3.2 Vorhergehende Schritte

In dieser Arbeit stehen die Segmentierung und Erkennung der handgeschriebenen Ziffern und Wörter im Vordergrund. Bevor diese Module zum Einsatz kommen, sind Verarbeitungsschritte erforderlich, um vom originalen Dokument zu einer digitalen Repräsentation zu gelangen, die als Grundlage der weiteren Verarbeitung geeignet ist.

Folgende Schritte sind hier zu nennen:

- Erzeugen eines digitalen Farbbildes
- Reduzierung zum Grauwertbild
- Erzeugen eines Binärbildes
- Erzeugen des Skeletts der Schriftobjekte
- Segmentierung der Zeilen

3.2.1 Digitalisieren

Es stehen mittlerweile unterschiedliche technische Lösungen zur Verfügung, um die Seiten eines Dokuments optisch abzutasten. Zu nennen sind die Digitalkamera, der Buchscanner und der Flachbettscanner. Jedes dieser Geräte kann für die Digitalisierung genutzt werden, wobei es unterschiedliche Vor- und Nachteile gibt.

Für den Vorgang der Digitalisierung – besonders mit Digitalkamera – ergeben sich potentielle Problemquellen, die beachtet werden müssen [12]:

- Ungleichmäßige Ausleuchtung
- Perspektivische Verzerrung
- Weitwinkel-Verzerrung
- Komplexer Hintergrund

- Fokussierung
- Intensität und Farbquantisierung
- Sensor Rauschen
- Kompression

Digitalkamera. Dank der rasanten technischen Entwicklung auf dem Gebiet der Digitalkameras sind CCD-Chips mit 5 Megapixel schon nahezu als Standard zu bezeichnen. Rein rechnerisch ist dies ausreichend. Um einen Eintrag auf einer A4-Seite mit 300 dpi zu digitalisieren, würde theoretisch eine Kamera mit 4,2 Megapixel ausreichen. Es muss dabei allerdings berücksichtigt werden, dass jegliche von der Kamera durchgeführte Verarbeitung und Aufbereitung der Daten die Bildauflösung reduzieren kann. Dies gilt insbesondere für die Verwendung von Kameras, die alle drei Farben mit einem Chip aufnehmen und anschließend eine Interpolation der Farbinfomationen durchführen. Letztendlich ist es eine Frage des Preises und der Zeit; selbst Kameras mit 11 Megapixel können schon erworben werden.

Buchscanner. Sehr gute Ergebnisse bei minimaler Materialbelastung verspricht die Verwendung eines Buchscanners. Diese Geräte sind speziell für die Digitalisierung von Büchern konzipiert und verfügen u. a. über Techniken zur Seitenwölbungskorrektur. Allerdings ist der finanzielle Aufwand relativ groß, sodass diese Variante nur für den kommerziellen Einsatz in Frage kommt.

Flachbettscanner. Auch Flachbettscanner liefern sehr gute Daten. Hier muss aber darauf geachtet werden, dass das Material nicht zu stark beansprucht wird. Der Einband eines alten Buches sollte nicht übermäßig belastet werden. Finanziell gesehen ist diese Variante am günstigsten. Auch die billigsten Scanner bieten heutzutage eine mehr als ausreichende Auflösung. Die möglicherweise verstärkt auftretenden Störungen wie perspektivische Verzerrung sind hier zu vernachlässigen.

Die in dieser Arbeit genutzten Daten wurden mittels Flachbettscanner digitalisiert. Dies setzte zwar ein sehr behutsames und damit zeitaufwendiges Vorgehen voraus, ergab aber sehr gute Resultate bei vergleichsweise geringen Kosten.

3.2.2 Vorverarbeitung

Ziel der Vorverarbeitung ist eine Repräsentation der Schriftobjekte in Form des Skeletts. Nach dem Digitalisieren liegen die Daten als 24-Bit-Farbbild vor. Nach der Auswahl eines Farbkanals (in der Regel ist dies der grüne oder blaue Kanal) wird eine Binarisierung durchgeführt, die einer Einteilung der Bildpunkte in Vorder- und Hintergrund entspricht. Anschließend erfolgt eine Skelettierung, deren Ergebnis die Grundlage der weiteren Verarbeitungsschritte darstellt.

(a) (b) (c)

Abbildung 3.1: Geräte, um Dokumente zu digitalisieren. (a) Digitalkamera (Canon EOS D60), (b) Aufsichtscanner (Zeutschel Omniscan 8000-3S), (c) Flachbettscanner (UMAX Astra1220).

Binarisieren

Zur Binarisierung wird auf dem Grauwertbild zunächst eine Hochpassfilterung mit einem Radius von ca. einem Millimeter durchgeführt. Anschließend wird mit einem konstanten Schwellwert binarisiert, der je nach Papier- und Tintenqualität bei 43 bis 45 % (Grauwert 111 bis 115) trennt.

Dieser Wert resultiert aus der Differenz zwischen der Helligkeit des Hintergrundes und der Helligkeit der schwach ausgeprägten Striche. Diese Differenz variierte über die unterschiedlichen Dokumentseiten nur sehr wenig und lag im Mittel bei 20 % (51 Grauwertstufen). D. h. durch eine Hochpassfilterung liegen die Werte des Hintergrunds in der Nachbarschaft einer schwachen Kante bei 60 % während die Werte der dunkleren Schriftobjekte höchstens bei 40 % liegen. Die Werte des Hintergrunds liegen bei 50 %, wenn sich keine Vordergrundobjekte in der Nähe befinden. Damit sollte der Schwellwert größer sein als 40 % und kleiner als 50 %.

In den alten Dokumenten treten des öfteren Verfärbungen des Papiers auf, die durch Verunreinigung, Stockflecken oder durch das Durchscheinen des rückseitigen Textes verursacht wurden. Die Helligkeitswerte dieser Störungen liegen in einem ähnlichen Bereich wie dünne oder schwach gezeichnete Striche des Textes. Da der Schwellwert so gewählt wurde, dass diese Striche während der Binarisierung nicht verloren gehen, wurden auch die Störungen als Teil des Textes erfasst.

Um die Zahl der falschen Vordergrundpixel zu reduzieren, wird der Umstand ausgenutzt, dass die Objektkanten dieser Störungen unschärfer sind. An Schriftobjekten treten höhere Frequenzen auf als an Störungen. Die gefundenen Vordergrundpixel werden dahingehend überprüft, ob sich in der Nähe ein Gradient mir einem hohen Betrag befindet. Ist dies nicht der Fall, werden diese Pixel dem Hintergrund zugeordnet.

Skelettieren

Für die strukturelle Analyse der Schrift sind es die Striche, die als Merkmal betrachtet werden. Diese Striche werden durch das Skelett besser repräsentiert als eine Binärbild-Repräsentation. Entscheidende Fragen lassen sich besser beantworten: Welche Lage und Länge hat der Strich? Wie ist seine Position relativ zu anderen Strichen?

Aus diesem Grund wird eine Skelettierung des zuvor gewonnenen Binärbilds durchgeführt. Hierbei wurde auf das Skelettierungsmodul der Software VECTORY der Firma Graphikon, Berlin [25] zurückgegriffen. Es erzeugt eine Kettencode-Repräsentation der Strichsegmente in einer Datenstruktur, die die topologischen Informationen repräsentieren. Während der Verarbeitung werden kleinere Störungen des Binärbilds toleriert, ohne ein zu starkes Ausfransen des Skeletts zu erzeugen.

3.2.3 Zeilensegmentierung

Für eine gut funktionierende Erkennung der Wörter eines Dokuments ist eine verlässliche Segmentierung von großer Bedeutung. Dies gilt sowohl für die zuvor beschriebene Segmentierung der Bildpunkte in Vorder- und Hintergrund als auch für die Segmentierung der Textobjekte eines Abschnitts oder einer Seite in Zeilen und die Segmentierung der Textobjekte einer Zeile in ihre Wörter.

Kann davon ausgegangen werden, dass die Zeilen gerade und parallel verlaufen, so bietet sich die Projektionsprofilmethode an. Für jede Zeile des Rasterbildes werden alle Vordergrundpixel [27] oder die Werte der Grauwertpixel [56] aufsummiert. Das daraus resultierende vertikale Profil wird nach Glättung zur Bestimmung der vertikalen Position der Zeilen und Zeilenzwischenräume genutzt, indem die lokalen Minima und Maxima erfasst werden. Verlaufen die Zeilen nicht gerade, liefert diese Methode keine verlässlichen Ergebnisse. Es ist erforderlich, ein angepasstes Verfahren wie die Zonenmethode zu nutzen [86]. Mehrere vertikal verlaufende Zonen ermöglichen die Verarbeitung gekrümmter oder geneigter Zeilen. Dennoch funktioniert dieser Ansatz nur für ordentlich geschriebene Zeilen, die gut voneinander getrennt sind. Dass diese Merkmale nicht auf die Schriften in Kirchenbüchern zutreffen, wurde bereits in Kapitel 2 erläutert. Aus diesem Grund ist die Anwendung eines dafür entwickelten Verfahrens zur Zeilensegmentierung erforderlich [17].

Die lokalen Extrema werden wie in dem Verfahren von Pu et al. [76] bestimmt und deren Anordnung untersucht. Allerdings wird der Umstand berücksichtigt, dass die vorliegenden Textzeilen nicht auf geraden Linien verlaufen müssen [50].

Der Algorithmus unterteilt sich in fünf Arbeitsschritte:

1. Finden der lokalen Minimum- und Maximumpunkte
2. Bilden der potentiellen Basisliniensegmente (pBLS)
3. Bewerten der pBLS-Alternativen und Entfernen der schlechten

4. Verbinden der Basisliniensegmente (BLS) zu den fertigen Basislinien
5. Finden der Mittellinien auf Grundlage der Basislinien
6. Zuordnung der Schriftobjekte zu den Zeilen

Die Zuordnung der Textobjekte zu den einzelnen Zeilen basiert auf der relativen Position. So wird ein Objekt einer Zeile zugeordnet, wenn es deren Basislinie schneidet. Durch Berührungen der Zeilen kann es vorkommen, dass sich ein Objekt über mehrere Zeilen erstreckt. Solche Objekte werden an geeigneten Stellen getrennt, sodass diese Mehrdeutigkeiten minimiert werden. Allerdings existieren nach wie vor Objekte, die nicht eindeutig einer Zeile zugeordnet werden können. Objekte, die im Zeilenzwischenraum liegen, werden beiden Zeilen zugeordnet und erhalten den Status „unsicher“. Diese Information kann in der späteren Erkennung genutzt werden (siehe dazu [17]).

Robustheit der Textlinienfindung

In jedem der oben angeführten Arbeitsschritte gibt es mehrere Parameter, die für ein zufriedenstellendes Arbeiten der Textlinienfindung korrekt eingestellt sein müssen. Um diese Werte für den Nutzer handhabbar zu machen, wurden diese Parameter in Abhängigkeiten der vorliegenden Schrifthöhe und -breite gesetzt. Zu Beginn des 7. Kapitels sind die durch Tests ermittelten Faktoren aufgeführt. Weiterhin konnte gezeigt werden, dass ein ungefährer Wert der Schriftgröße für ein Funktionieren des Verfahrens ausreichend ist.

3.3 Wortsegmentierung

Um ein einzelnes Wort oder eine einzelne Ziffer in einem geschriebenen Text automatisch erkennen zu können, müssen die Grenzen dieses Objektes gefunden und die zugehörigen Textteile ermittelt werden. Aufgrund der besonderen Merkmale alter Aufzeichnungen ist ein Verfahren zur Wortsegmentierung nicht anwendbar, dass nur auf geometrischen Merkmalen wie dem Abstand von Textobjekten basiert (siehe Abbildung 3.2). Es muss zusätzliches Wissen über den Aufbau des Textes genutzt werden. Daher ist es erforderlich, das Problem der Wortsegmentierung und -erkennung einzugrenzen. In Kapitel 4 wird ein Algorithmus vorgestellt, der exemplarisch für die Datumseinträge mehrere Hypothesen für eine Segmentierung in die Bestandteile eines Datums erzeugt. Dabei werden zwei Informationsquellen kombiniert: Erstens führt das A-priori-Wissen über die mögliche Struktur der Wortfolge dazu, dass Grenzen zwischen Worten oder Ziffern nur an bestimmten Stellen wahrscheinlich sind. Zweitens gibt es bestimmte geometrische Merkmale von Textteilen, die eine Aussage über die Wahrscheinlichkeit einer Wortgrenze an einer bestimmten Stelle ermöglicht. Das Ergebnis sind Hypothesen über mögliche Positionen von Wortgrenzen, die es ermöglichen, mit wenigen Versuchen das korrekte Ergebnis zu finden.

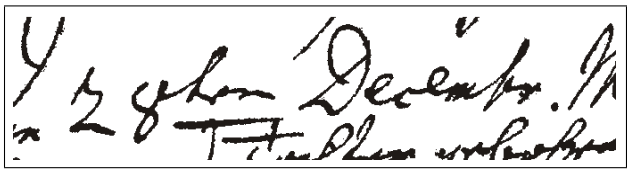


Abbildung 3.2: Beispiel einer Berührung benachbarter Worte bzw. Ziffern.

3.3.1 Ansätze zur Segmentierung handgeschriebener Wörter

Es gibt unterschiedliche Verfahren, die eine handgeschriebene Zeile in ihre Worte zerlegt. So existieren Verfahren, die die klassische Reihenfolge Segmentierung-Erkennung umkehren und eine Wortsegmentierung durchführen, die auf den zuvor erkannten einzelnen Zeichen beruht [35]. Dies kann jedoch nur in solchen Fällen angewandt werden, bei denen die Situation der Schrift so günstige Voraussetzungen schafft, dass eine Identifizierung der einzelnen Zeichen ermöglicht wird – wie z. B. in sorgfältig ausgefüllten Formularen. Da dies bei den hier vorliegenden Dokumenten nicht der Fall ist, muss auf anderem Wege eine Segmentierung der Zeilen in Worte erfolgen.

In den meisten Verfahren zur Wortsegmentierung werden Abstände zwischen den Schriftobjekten analysiert, um so Aussagen über Wortgrenzen machen zu können [42, 56, 58, 73]. Bei Seni et al. [84] werden horizontale Entfernungen (run-length) und euklidische Distanzen bestimmt, die besser geeignet sind als die simple Abstandsmessung zwischen den „Bounding Boxes“. Im Gegensatz dazu wird bei Mahadevan et al. [54] der Abstand zwischen den konvexen Hüllen betrachtet. Diese Lücken werden nach ihrer Größe sortiert, um anschließend Hypothesen für die so entstehende Wortfolge zu bilden. Kim et al. [40] nutzen ein neuronales Netz (NN), um potentielle Wortgrenzen in Addressangaben zu finden. Neben der Abfolge von großen und kleinen Buchstaben dient vor allen die Größe einer Lücke als Basis. Bei Tomai et al. [91] werden zwar auch nur die Abstände zwischen den Schriftobjekten betrachtet, allerdings werden hier mehrere Hypothesen über die Anzahl und Position der Wortgrenzen erzeugt, um erst später durch das Auswerten mehrerer Hypothesen eine Entscheidung für die Zusammensetzung einer Zeile zu treffen.

El-Yacoubi et al. [13] finden und erkennen Straßennamen in Addressangaben, indem ein entsprechend angepasstes Hidden Markov Model (HMM) verwendet wird. Eine explizite Segmentierung erfolgt nicht. Eine erfolgreiche Erkennung hat automatisch die Segmentierung des Straßennamen zur Folge.

Morita et al. [67] nutzen einen HMM-Ansatz, um das Datum auf Schecks in seine Bestandteile zu zerlegen. Ein zu segmentierender Satz besteht hier aus dem Namen der Stadt, dem Tag, dem Monat und dem Jahr. Es ist also von der Aufgabenstellung her vergleichbar mit dem hier vorliegenden Problem. Da es sich hier um Schecks mit Feldern für die einzelnen Teile eines Eintrags handelt, ist in der Regel viel Raum zwischen diesen Teilen. Es kommt nicht zu Berührungen. Dennoch könnte der HMM-Ansatz auch für die Segmentierung von Worten in historischen Dokumenten anwendbar

sein, wenn es gelingt, das A-priori-Wissen über die Struktur des zu erkennenden Textes in das System einzubringen.

Das Verfahren, das von Xu et al. [102] vorgestellt wird, führt ebenfalls eine Segmentierung und Erkennung des Datums auf Schecks durch. Zwar ist die Varianz der Datumsform der hier betrachteten kanadischen Schecks größer, es werden jedoch zusätzliche Merkmale zur Segmentierung der Datumsbestandteile betrachtet. In 80 % der Fälle existiert ein Vordruck des Jahrhunderts „19“ bzw. „20“ auf dem Dokument (siehe Abbildung 3.3). Desweiteren wird nach Trennzeichen wie Kommata, Bindestriche oder größeren Lücken gesucht. Daten mit sich berührenden Komponenten können von dem Verfahren nicht verarbeitet werden. Getestet wurde auf 1219 englischen und 2180 französischen Daten. Erzielt wurde eine korrekte Segmentierungsrate von 90,4 % (englisch) bzw. 82,94 % (französisch).

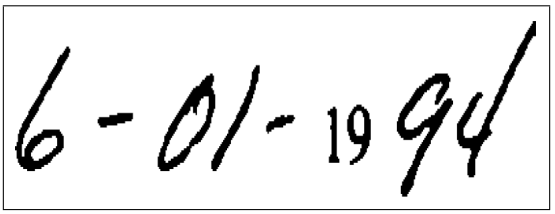


Abbildung 3.3: Datum eines kanadischen Schecks mit gut separierten Bestandteilen [102].

3.3.2 Hypothesen durch A-priori-Wissen

Die oben erwähnten Verfahren zur Wortsegmentierung befassen sich in der Regel mit Schriften, die nicht so problematisch sind wie die der Kirchenbücher. Um bei dieser Schrift Grenzen zwischen Wörtern finden zu können, wird in dem in Kapitel 4 beschriebenen Verfahren zusätzliches Wissen über die Struktur der vorliegenden Wörter in das System eingebracht. Sowohl die Wortsegmentierung als auch die Erkennung wird in dieser Arbeit am Beispiel des Datums eines Kirchenbucheintrags gezeigt. Hierfür existiert das Wissen, dass die Angabe des Datums nur aus einer kleinen Zahl von Objekten besteht und diese nur in bestimmten Konstellationen auftreten können.

In der momentanen Umsetzung ist es erforderlich, dass der Nutzer die Position des Datums im Eintrag manuell markiert. Je nach Schreiber wurde dieses Datum entweder am Anfang oder am Ende des Eintrags gesetzt. So sollte es möglich sein, diesen manuellen Schritt zu automatisieren.

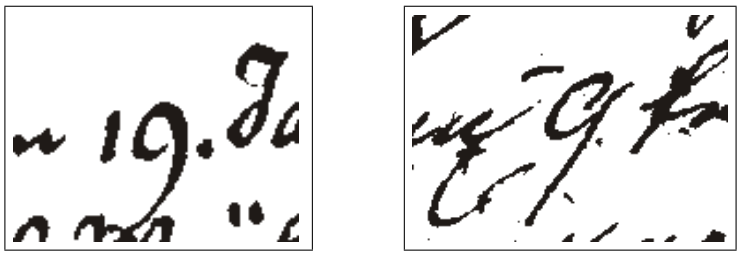
In einem ersten Schritt werden innerhalb des zu untersuchenden Textes Positionen für potentielle Wortgrenzen festgelegt. In erster Linie befinden sich diese an Lücken zwischen Schriftobjekten. Da bei der hier betrachteten Schrift nicht ausgeschlossen werden kann, dass sich benachbarte Wörter berühren, wird ein zweiter Typ von potentiellen Wortgrenzen an Stellen erzeugt, an denen der obere Verlauf der Hülle eines Objektes im Zeilenhauptraum ein lokales Minimum aufweist – die Position

eines Tals der oberen Seite. Dadurch werden auch Wortgrenzen erfasst, die durch eine Berührung zweier benachbarter Wörter gekennzeichnet sind.

Mit Hilfe dieser Menge potentieller Wortgrenzen werden anschließend Hypothesen über die Aufteilung des zu erkennenden Zeilenabschnitts erzeugt. Mit anderen Worten: Es existiert an dieser Stelle keine eindeutige Lösung für die Position der Wortgrenzen, sondern eine Liste von Hypothesen, von denen die besten zur weiteren Erkennung betrachtet werden. Damit wird weiter die erfolgreiche Strategie verfolgt, mehrere wahrscheinliche Interpretationen der Daten zuzulassen. Im weiteren Verlauf der Verarbeitung dient die Analyse neuer Daten dazu, Hypothesen zu stärken oder zu schwächen. Eine vorschnelle Entscheidung für die zu einem Zeitpunkt wahrscheinlichste Lösung wird vermieden.

3.4 Ziffernerkennung

Wie bereits in Abschnitt 2.4.3 erläutert, hat sich das Aussehen der arabischen Ziffern in den letzten Jahrhunderten nicht wesentlich geändert. Dennoch gibt es Abweichungen in der Strichabfolge, durch die eine Ziffer geschrieben wird (siehe Abbildung 3.4). Diese Unterschiede sind für eine stabile Erkennung von Bedeutung.



(a) Beispiel von 1773. (b) Beispiel von 1813.

Abbildung 3.4: Eine Ziffer kann auf unterschiedliche Art und Weise geschrieben werden.

3.4.1 Statistischer vs. Struktureller Ansatz

Grundsätzlich können die Ansätze, nach denen die Erkennung von Ziffern, Zeichen oder Worten erfolgt, in zwei Gruppen unterteilt werden: statistisch und strukturell [24]. Erkennungssysteme arbeiten entweder nach einem dieser Ansätze oder basieren auf einer Kombination beider. Jeder Ansatz bietet Vor- und Nachteile, die im Folgenden erläutert werden.

Statistischer Ansatz

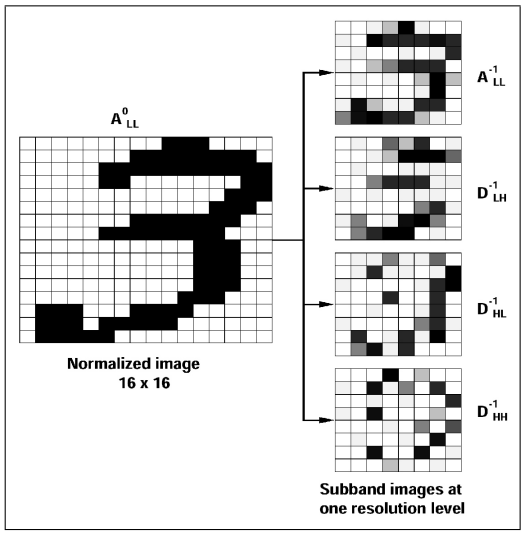
Statistische Verfahren zur Ziffern- oder Zeichenerkennung bieten bei gut segmentierten Kandidaten den Vorteil einer sehr guten Erkennungsrate bei vergleichsweise einfach strukturierten Merkmalen. Es liegt in der Natur der verwendeten Klassifizierungsverfahren, dass hier Merkmalsvektoren gebildet werden, deren Merkmalsraum eine festgelegte Zahl von Dimensionen besitzt.

Bei der Auswahl der Merkmale trifft man bei statistischen Verfahren auf sehr vielseitige Varianten. Diese reichen von der naheliegenden Analyse der Verteilung der schwarzen Bildpunkte im normalisierten Kandidatenbild [11] bis hin zu komplexeren Merkmalen wie Ecken und Kanten oder Analysen des Frequenzraums.

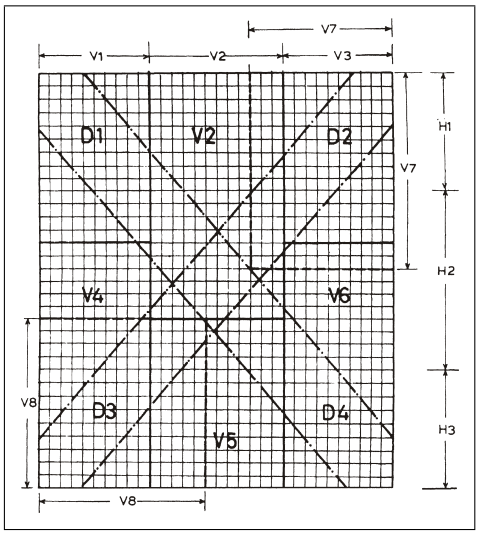
So wird in [8] die räumliche Verteilung von hohen und niedrigen Frequenzen auf der Basis einer 2D Wavelet Transformation von vier Bildern der Größe 8 x 8 analysiert. Diese wurden zuvor aus einem 16 x 16 großen normierten Binärbild der Ziffer erzeugt (Abbildung 3.5(a)).

In anderen Verfahren wird die Anordnung der Pixel des Binärbildes in konzentrisch angeordneten Ringen [29] oder in horizontalen, vertikalen und diagonalen Regionen – so genannte „Bar Masks“ – herangezogen [79] (siehe Abbildung 3.5(b)).

Durch die Suche nach Kanten und Ecken im Grauwertbild werden in dem Verfahren von Teow et al. [90] 32 9 x 9 Matrizen als Merkmalsmenge erzeugt. Es wird hier also ein sehr hochdimensionaler Merkmalsraum erzeugt.



(a) Merkmalsextraktion durch Analyse der Verteilung von hoch- und niederfrequenten Stellen im Bild mittels Wavelet Transformation [8].



(b) Merkmalsextraktion durch Analyse der Anordnung der Objektpunkte in bestimmten Regionen, so genannten „Bar Masks“ [79].

Abbildung 3.5: Bei statistischen Erkennungsverfahren existieren sehr vielseitige Varianten der Merkmalsextraktion.

Weitere Möglichkeiten, die Form eines zu erkennenden Zeichens zu repräsentieren, wie beispielsweise die Distanz-Transformationen zwischen Vordergrund- und Hintergrundpixeln, werden in [70] erläutert.

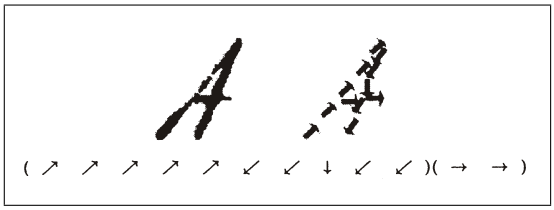
All diese Daten werden als Eingabe für verschiedene Lern- bzw. Klassifikationsverfahren genutzt wie Single-Layer- und Multi-Layer-Perceptrons oder Support-Vector-Machines.

Ein großer Nachteil des statistischen Ansatzes ist die Notwendigkeit eines großen Datensatzes, um das System zu trainieren. Man erhält einen Erkenner, der für eine Schriftart sehr gut funktioniert. Ändert sich die Charakteristik, dann muss das System auf die neuen Daten angepasst werden. D. h. es ist ein neuer großer Trainingsdatensatz erforderlich. Das „Black-Box"-Verhalten eines solchen Systems erschwert das Ergründen von Fehlleistungen. Zusätzliches Wissen über Charakteristiken der Schrift kann bei diesem Ansatz nicht genutzt werden.

Struktureller Ansatz

Beim Schreiben eines Wortes oder einer Ziffer werden bewusst Striche in einer bestimmten Form und Anordnung erzeugt. Und auch beim Lesen ist es diese Anordnung als primäres Merkmal, die den Menschen – oftmals unbewusst – eine bestimmte Hypothese für die eine oder andere Ziffern besonders hoch bewerten lässt. Merkmale wie Projektionsprofile der Kontur [28], sind hingegen sekundäre Merkmale, die durch die Anordnung der Striche gebildet werden.

Bei strukturellen Verfahren ist es häufig der Verlauf der Striche, der direkt als Merkmal analysiert wird. Bei Cha et al. [7] wird dies anhand der Kontur durchgeführt. Ein Strich wird dabei in mehrere



(a)

	1st part	2nd part
1	↑↗↑↗→↘↓↘↓	↘↗
2	↗↗↗↗↗↙↙↓↙↙	→→
3	↗↑↑↗→↘↓↓↙	→↗
4	↗↑↑↗→↘↓↓↙	→↗
5	↗↗↑↗↗↘↓↘↓↓	→→↗
6	↓↓↓↑↑↑↗↗↘↓↓↘↓	↗→
7	↓↓↓↑↑↗↗↗↘↓↓↓↓	→→
8	↗↗↗↗↗↗↓↓↓↙	→→→
9	↗↗↗↗↗↗↓↓↙↓↙	→→→
10	↗↗↗↗↗↓↙↓↓↓	→→→
11	↙↙↙↙↗↗↗↗↗↗↓↙↓↙↙	→→→
12	↙↙↙↙↙↗↗↗↗↗↗↓↙↙↙↙	→→↗
13	↙↙↙↙↗↗↗↗↗↓↙↙↙↙	→↗
14	↙↙↙↙↗↗↗↗↗↓↙↙↙	→↗
15	↗↗↗↗↗↘↓↓↓↙	→↗
16	↗↗↗↗↗↓↓↓↓↓	↗→
17	↓↙↓↙↗↑↗↑↗↘↓↓↓↓	↗↗
18	↓↙↙↗↗↗↗↗↘↓↓↓↓	→→→
19	↓↙↙↗↗↑↗↗↓↓↓↓↘	→↗
20	↙↙↓↙↗↗↑↗↗↓↓↓↓↓	→↗

(b)

Abbildung 3.6: Beispiel für die Extraktion und Analyse struktureller Merkmale [7].

kleine Strichsequenzen mit konstanter Länge zerlegt. Die Abfolge der unterschiedlichen Richtungen wird dann in einer Distanzmatrix verglichen. Hier besteht jedoch das Problem, dass eine hohe Anfälligkeit gegenüber Lücken im Strichverlauf und Berührungen mit störenden Strichen besteht. Das Skelett der Schriftobjekte bietet bessere Möglichkeiten, den Verlauf der Striche zu rekonstruieren. Dies wird in [7] für eine Online-Erkennung genutzt (siehe Abbildung 3.6). Gleichzeitig greift man aber – wie bei Online-Verfahren üblich – auf den Pfad des Schreibgerätes zurück.

Auch im Bereich der Erkennung von gedruckten Zeichen gibt es Verfahren, die direkt die Form eines Zeichens analysieren und bewerten. Rocha et al. [80] versuchen, den Entstehungsprozess des Zeichens umzukehren und aus der Position der Bildpunkte Linien zu rekonstruieren, deren Anordnung mit Prototypen verglichen und bewertet wird.

Khan et al. [37, 38] beschreiben ein Verfahren zur Erkennung handgeschriebener Buchstaben und Ziffern. Es werden Protoypen genutzt, die auf dem geometrischen Modell von handgeschriebenen Zeichen basieren (siehe Abbildung 3.7). Die Prototypen setzen sich aus Primitiven in Form von Geraden, viertel Kreisen, halben Kreisen und kompletten Kreisen zusammen. Das Verfahren zeichnet sich durch eine Robustheit gegenüber störenden Linien aus. Nachteilig ist die Verwendung der vier festgelegten Primitivformen. Foggia et al. [22] beschreiben einen kombinierten Ansatz, bei dem strukturelle Primitive genutzt werden, deren Abweichung statistisch ausgewertet wird. Primitive sind hier Kreisteile mit einem frei definierbaren Winkel. Es wird ein Multi-Layer Perceptron verwendet, das geometrische Momente der strukturellen Repräsentation bewertet. Dieser Klassifizierungstechnik ist auf Grund des notwendigen Trainings und des „Black-Box“-Verhaltens für die strukturelle Analyse und die Adaptation des Verfahrens nicht geeignet.

Strukturelle Verfahren setzen bei der Generierung der Prototypen mehr manuelle Arbeit voraus. Auch liegt die Erkennungsrate eines strukturellen Verfahrens in der Regel etwas unter der eines

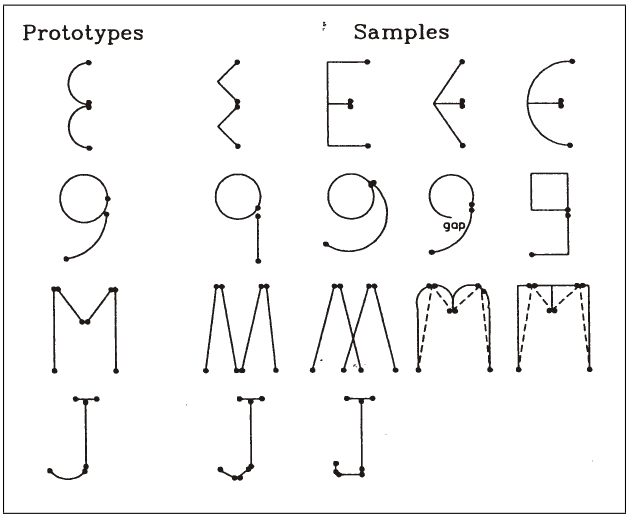


Abbildung 3.7: Bei einem strukturellen Verfahren wird die Anordnung der Striche als Merkmal untersucht [38].

optimal trainierten statistischen. Ein strukturelles Verfahren kann jedoch ohne jegliches Training eingesetzt werden. Dies ist vor allem dann ein Vorteil, wenn für ein gegebenes Problem kein großer Trainingsdatensatz besteht. Da das Vorhandensein und die Position von Strichen analysiert und bewertet wird, kann dieses transparente Verhalten genutzt werden, um Anpassungen an andere Schriftstile vorzunehmen.

Kombination beider Ansätze

Um einerseits die Struktur eines Musters explizit zu beschreiben und andererseits die Informationen in großen Datensätzen nutzen zu können, gibt es Verfahren, die sowohl den strukturellen als auch den statistischen Ansatz zur Erkennung von Ziffern vereinen. So findet bei Hu et al. [31] zunächst eine Vorklassifizierung auf der Basis von strukturellen Eigenschaften wie die Zahl von Linienenden oder Schleifen statt, um anschließend diese Superklassen, die noch mehrere Ziffern vereinen, durch trainierte neuronale Netze in die finalen Klassen der einzelnen Ziffern zu zulegen. Der Nachteil ist hier die hohe Zahl an Trainingsdaten. Selbst bei einer Größe von ca. 20 000 Ziffern verteilten sich die Kandidaten sehr unterschiedlich auf die 26 Superklassen. In einem Fall entfiel auf eine Superklasse nur ein Kandidat.

Auch bei Lee et al. [47] findet eine Verarbeitung in zwei Schritten statt. In einem ersten Schritt werden die Ziffern anhand von strukturellen Merkmalen wie Zahl der Schnittpunkte an einer senkrechten Scan-Linie vorklassifiziert. In einem zweiten Schritt erfolgt eine Klassifizierung mittels neuronalen Netzes.

Eine andere Art der Kombination des statistischen und strukturellen Ansatzes stellen Verfahren dar, die strukturelle Merkmale statistisch auswerten. Hierbei besteht das Problem, dass die strukturellen Merkmale, die in unterschiedlicher Anzahl auftreten, in einen Merkmalsvektor fester Länge überführt werden müssen [30, 22].

3.4.2 Der eigene strukturelle Ansatz

Es soll an dieser Stelle nicht die maximal erreichbare Erkennungsrate einer Ziffer im Vordergrund stehen. Ein entsprechend trainiertes statistisches Verfahren erreicht vermutlich bessere Werte als das hier präsentierte. Das Aussehen der arabischen Ziffern variiert nicht so systematisch, als dass durch das Anpassen bestimmter Parameter eine signifikante Verbesserung der Erkennung zu erwarten wäre. Es wurde ein struktureller Ansatz verfolgt, um Erkenntnisse zu gewinnen, die für die Erkennung der Monatsnamen von Bedeutung sind. Grundlegend konnte hier der Zusammenhang zwischen dem Begriff der Ähnlichkeit und den einzelnen Transformationen erforscht werden. Über das Erfassen und Vergleichen von Abweichungen der vorliegenden Striche von den erwarteten kann somit eine

Bewertung für die einzelnen Einträge des Lexikons vorgenommen werden – sowohl für die Ziffern- als auch für die Worterkennung.

Das hier vorgestellte Erkennungsverfahren ist an den strukturellen Ansatz von Khan et al. [37, 38] angelehnt. Die Erkennung einer Ziffer erfolgt durch die Verwendung von Prototypen. Im Prototypen wird die Anordnung von Strichen repräsentiert, die für das zu erkennende Objekt markant und entscheidend sind. Im Bezug auf kleinere Objekte wie Ziffern sind dies in der Regel sämtliche Striche, während es bei Wörtern zumindest die Striche markanter Stellen sind.

Während des Matchings zwischen einem Kandidaten und einem Prototypen wird untersucht, inwieweit diese Merkmale im Kandidaten gefunden werden können. Das Verfahren unterteilt sich in zwei Teilprozesse: das Erzeugen der approximativen Repräsentation eines Ziffernkandidaten und das Matching mit den Prototypen der Ziffern.

Ein Prototyp besteht aus der Anordnung von Primitiven (Kreisbögen). Während der Erkennung wird für jedes Primitiv des Prototypen das am besten passende Primitive der approximativen Repräsentation des Kandidaten gesucht. Der Grundgedanke für die Erkennung einer Ziffer besteht darin, dass die notwendigen Transformationen der Primitive des richtigen Prototypen geringer sind als für die Primitive des falschen Prototypen.

3.5 Worterkennung

Auch wenn die komplette Erkennung des Textes einer Kichenbuchseite noch nicht realisierbar ist, kann die Erkennung einzelner Wörter einerseits für den Historiker oder Genealogen recht hilfreich sein und andererseits wichtige Erkenntnisse für die Erkennung solcher Schrift liefern.

Die Besonderheiten alter Schrift treten während der Erkennung von Ziffern noch nicht stark in den Vordergrund. Sollen jedoch in alter Handschrift geschriebene Wörter erkannt werden, sind die meisten existierenden Verfahren zur Worterkennung nicht oder schlecht geeignet. Der Grund dafür liegt u. a. in der engen Schreibweise der Zeilen und Wörter aber auch in dem bereits erwähnten Fehlen großer Trainingsdatensätze.

3.5.1 Verfahren zur Worterkennung

Mit dem Menschen als Vorbild sind die Ziele der Forschung hoch gesteckt, wenn es darum geht, handgeschriebene Texte automatisch erkennen zu können. Es wurden vielfältige Techniken entwickelt, um das Bild eines geschriebenen Wortes als dieses Wort zu erkennen.

Das Feld der Worterkenner lässt sich ebenso wie das der Ziffernerkenner in strukturell und statistisch einteilen. Darüber hinaus ergeben sich weitere Klassifizierungsmerkmale, die die Art und Weise

der Auswertung der Merkmale betrifft. Grundsätzlich besteht eine Zweiteilung der Verfahren basierend auf der Existenz eines Segmentierungsschrittes. Daraus ergibt sich einerseits die Klasse der auf Segmentierung basierenden Verfahren und andererseits die Klasse der holistischen Verfahren.

Segmentbasierte Erkennungsverfahren

Die Merkmale eines Wortes entstehen durch die Aneinanderreihung von Buchstaben. Daher liegt es nahe, die Merkmale der Buchstaben zu trainieren und während der Erkennung nach diesen Merkmalen zu suchen. Dies ist auch erforderlich, wenn die Zahl der Einträge des Lexikons zu groß wird. Das dabei entstehende Problem besteht in dem Finden der korrekten Grenzen zwischen den Buchstaben. Anders als bei Druckschrift können hier nicht die Lücken zwischen den Buchstaben als verlässliches Merkmal herangezogen werden. Die Grenzen zwischen den Buchstaben können nur dann sicher und fehlerfrei gefunden werden, wenn diese Buchstaben bekannt sind. Auf den Punkt gebracht wird dieses Dilemma durch das Paradoxon von Sayre [81]:

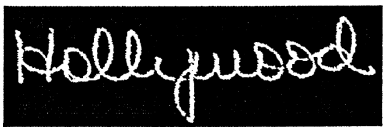
> *"To recognize a letter, one must know where it starts and where it ends, to isolate a letter, one must recognize it first."*

Ein typisches Vorgehen besteht in der Übersegmentierung eines Wortes in viele kleine Abschnitte – genannt Grapheme. Durch das Zusammenfassen benachbarter Grapheme wird versucht, die Buchstaben des Wortes zu bilden (siehe Abbildung 3.8). Das Problem der richtigen Kombination der Grapheme wird mittels dynamischer Programmierung [5, 16, 2] oder durch Hidden Markov Models gelöst [14, 23]. Es existieren auch Kombinationen bei denen beispielsweise ein neuronales Netz Wahrscheinlichkeiten für die einzelnen Abschnitte bestimmt und diese anschließend mittels HMM kombiniert werden [85].

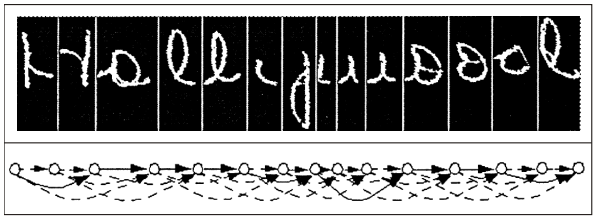
Ein weiteres Beispiel für ein auf Segmentierung basierendes Verfahren wird von Tay et al. [88] beschrieben. Für die Erkennung von auf Schecks befindlichen Wörtern wird hier eine Kombination aus Neuronalem Netz und Hidden Markov Model genutzt. Mit Hilfe von existierenden Datenbasen mit einem Umfang von mehr als 36.000 isolierten Wörtern konnten die Klassifikatoren erfolgreich trainiert und getestet werden.

Im Bezug auf die Verwendung von Hidden Markov Modells ergibt sich ein weiterer Ansatz, indem die Segmentierung implizit während der Erkennung stattfindet. Dabei kommen sogenannte *continuous density HMMs* zum Einsatz, die Merkmale eines jeden Segments eines Wortes bewerten [64]. Dazu müssen auch hier zuvor Modelle für die einzelnen Buchstaben trainiert werden.

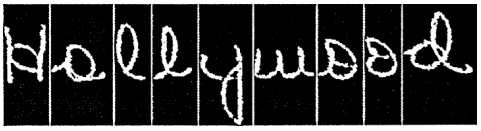
Das Verfahren von Xiao et al. [101] analysiert ein Wort hinsichtlich seiner nach oben oder nach unten weisenden Täler sowie Schleifen. Anhand von Regeln wird das Wort in so genannte Subkomponenten von Zeichen aufgeteilt. In einem nächsten Schritt werden diese zu Zeichen zusammengefügt.



(a) Originalbild des Wortes.



(b) Grapheme und Segmentierungsgraph.



(c) Segmente und korrekter Graph der Segmentierung.

Abbildung 3.8: Übersegmentierung gefolgt von dynamischer Programmierung [104].

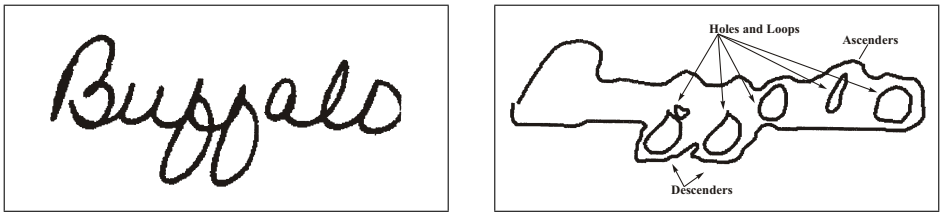
Dieser Algorithmus ist nach eigenen Aussagen für schwierige Verhältnisse wie unsauber geschriebene Schrift oder sich berührende Zeichen (jenseits der Ligaturen) nicht geeignet.

Ein Verfahren, das speziell die Erkennung von Monatsnamen betrifft, wird von Morita et al. präsentiert [66, 65]. Zuvor isolierte Monatsnamen auf Bankschecks sind der Gegenstand der Erkennung. Der vorgestellt Ansatz beruht auf der Verwendung eines HMMs mit vorhergehender Segmentierung. Der verwendete Zeichenerkenner wurde mit Hilfe eines Datensatzes von ca. 10.200 Wörtern trainiert. Neben stark gestörten Wörtern bereiteten Wörter mit fehlenden oder falschen Ober- oder Unterlängen dem Verfahren Probleme.

Ein ebenfalls für die Erkennung der Monatsnamen auf Bankschecks entwickeltes Verfahren wird von Xu et al. vorgestellt [102]. Dabei wird ein auf Graphemen basierendes HMM mit zwei Multi-Layer Perceptrons kombiniert. Das Verfahren wurde auf 2063 Daten getestet. Erreicht wurde eine Erkennungsrate von 87,06 %.

Holistische Erkennungsverfahren

Das Zerlegen eines Wortes ist aufwändig und für dessen Erkennung nicht immer notwendig. Wird ein Wort als Einheit betrachtet und werden daraus geeignete Merkmale extrahiert, so spricht man



(a) (b)

Abbildung 3.9: Die holistische Worterkennung nutzt Merkmale des gesamten Wortes wie Wortlänge, Ober- und Unterlängen und Schleifen [51].

von einem holistischen Ansatz. Wichtige Merkmale sind die Länge eines Wortes, Schleifen sowie die Ober- und Unterlängen (siehe Abbildung 3.9).

Parisse [72] nutzt diese Daten indirekt, indem die obere und untere Kontur extrahiert werden. Die Erkennung erfolgt durch eine Suche nach dem ähnlichsten gespeicherten Profil.

Das Besondere and der Arbeit von Cai et al. [6] ist eine geringe Zahl von Trainingsdaten. Das Verfahren zur schreiberunabhängigen Erkennung, das an Städte- und Straßennamen getestet wurde, benötigte lediglich 3 bis 16 Beispieldaten. Allerdings erreicht das Verfahren bei einer Lexikongröße von 14 Wörtern eine Erkennungsrate von 64,6 % und liefert keine weiteren Erkenntnisse über Merkmale der Schrift.

Wahrnehmungsorientierte Erkennungsverfahren werden in einigen Arbeiten als eigene Klasse angeführt [93, 87]. Sie können prinzipiell jedoch zu den holistischen Verfahren gezählt werden, da das menschliche Lesen von Wörtern auf holistische Weise erfolgt. Die Ansätze basieren auf der Beobachtung, dass ein menschlicher Leser ein Wort nicht von links nach rechts abtastet, um die Abfolge von Merkmalen zu erfassen, sondern die Form des Wortes als Gesamtheit betrachtet und seine Entscheidung von bestimmten, markanten Schlüsselzeichen abhängig macht [89]. Schomaker et al. [83] untersuchten mit Hilfe von Testpersonen, welche Stellen in einem Wort für dessen Erkennung von höherer Bedeutung sind. Dabei zeigte sich, dass Zeichen mit Ober- oder Unterlängen wichtiger sind als die übrigen Zeichen. Desweiteren wurde festgestellt, dass dies auch für das erste und das letzte Zeichen eines Wortes gilt.

3.5.2 Erkennung historischer Texte

Verglichen mit dem gesamten Feld der Schrifterkennung existieren momentan relativ wenig Arbeiten, die sich speziell mit der Erkennung von Schrift in alten Dokumenten beschäftigen.

Eine Alternative zur klassischen Worterkennung mittels Klassifikator wird als „word spotting“ bezeichnet [55, 77, 78]. Sie wurde für historische Texte entwickelt und fasst mehrere Instanzen eines

Wortes zu einer Gruppe zusammen. Während der Erkennung wird die beste Gruppe gesucht, indem Merkmale wie vertikale Projektionsprofile verglichen werden. Ebenso wie beim Verfahren von Lavrenko et al. [46] wird das Problem auf einen Schreiber eingeschränkt. Getestet wurden die Verfahren an Schriften von George Washington (siehe Abbildung 3.10). Ändert sich der Schreiber, muss das System neu trainiert werden. Desweiteren zeichnen sich diese Dokumente durch eine saubere Schreibweise aus. Der Zeilenverlauf ist gerade. Zeilen und Worte sind in der Regel gut voneinander getrennt.

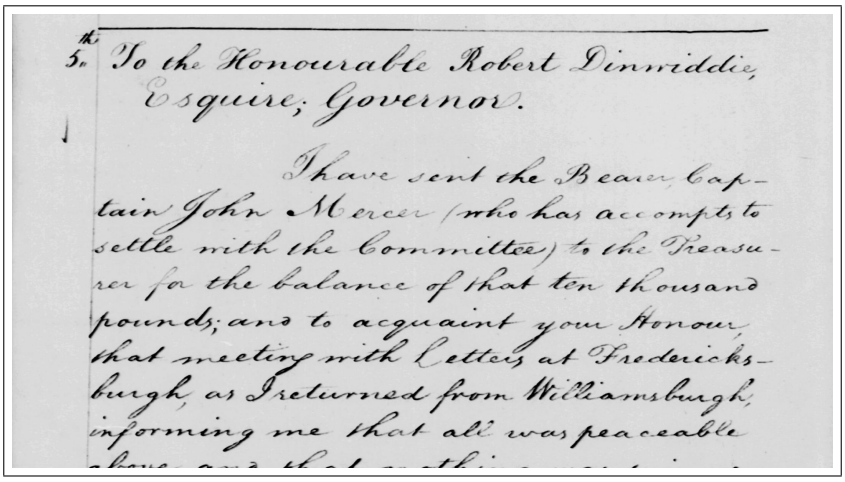


Abbildung 3.10: Beispiel der Schrift aus Briefen von George Washington [46]. Vor allem das Vorhandensein von nur einem Schreiber verringert die Variabilität und damit das Problem der Erkennung.

Kolcz et al. [44] bezeichnen mit „word spotting“ die Suche nach einem bestimmten Wort in Seiten eines historischen Dokuments. Dieser Zwei-Klassen-Ansatz (gesuchtes Wort / nicht gesuchtes Wort) wurde auf Dokumenten des „Archive of the Indies“ getestet, der Briefkorrespondenz zwischen der Spanischen Krone und den Kolonien der Neuen Welt vom 15. bis zum 19. Jahrhundert. Dabei lehnt sich diese Arbeit an der von Keaton et al. an [36]. Auch hier bilden diese Dokumente den Gegenstand der Untersuchungen. In beiden Arbeiten wird das Profil der oberen und unteren Hülle eines Wortes als Merkmal zu Klassifizierung herangezogen.

Ähnliche Aussagen wie über die Dokumente von George Washington lassen sich über Briefe von Thomas Jefferson machen, die Gegenstand der Untersuchungen von Tomai et al. [91] sind. Das dort vorgestellte Verfahren hat zum Ziel, den kompletten Text einer Briefseite zu erkennen. Mehrere Hypothesen über die Wortsegmentierung einer Zeile werden mittels Worterkennung [39] getestet. Neben dem Vorhandensein von lediglich einem Schreiber ist die Qualität der Schrift ein entscheidender Unterschied zu Dokumenten wie Kirchenbücher. Die Briefe von George Washington oder Thomas Jefferson wurden sorgfältiger geschrieben als Kirchenbücher hiesiger Gemeinden. Vor allem der kostbare Platz auf dem Papier wirkt sich stark auf die auftretenden Störungen im Schriftbild aus.

3.5.3 Adaptierbare Worterkenner

Sind die Bedingungen schwierig, ist eine Anpassung des Verfahrens zur Worterkennung notwendig, um trotz auftretender Artefakte eine Erkennung zu ermöglichen. Daher ist ein Verfahren notwendig, das die Möglichkeit zur Anpassung an eine Schrift bietet. Wie bereits in Kapitel 2 beschrieben, ist die Schrift eines Dokuments durch Merkmale gekennzeichnet, die zum Teil durch eine Einordnung in Schriftklasse und Schriftform erfasst werden können.

Auf dem Gebiet der On-Line-Handschrifterkennung gibt es Bestrebungen, eine Anpassung an einen Schreiber durchzuführen. So untersuchten Vuori et al. [96] vier Strategien zu Schreiberanpassung. U. a. wird hier die Anpassung eines existierenden Prototypen durch Umformung beschrieben, wenn eine ausreichend hohe Ähnlichkeit zwischen dem Prototypen und dem Kandidaten vorliegt. Andernfalls wird ein neuer Prototyp für diesen Kandidaten erzeugt. Es findet keine Klassifizierung bestimmter Schriftstile statt, sodass a priori keine Anpassung des Verfahrens erfolgen kann. Es wird, wie auch in anderen Arbeiten, ein statistischer Ansatz verfolgt, bei dem HMMs [3, 4, 94] oder Neuronale Netze [60] zum Einsatz kommen.

Schon seit längerer Zeit gibt es Bestrebungen, die Charakteristiken einer Handschrift zu untersuchen und in Stil-Familien zu kategorisieren [10, 97]. Es sind Merkmale wie Strichdichte im Zeilenhauptraum oder die Richtungen der Striche – Merkmale, die nicht über die Entstehung sondern aus dem Bild der Schrift gewonnen werden.

Das in dieser Arbeit vorgestellte Verfahren basiert ebenso wie die Ziffernerkennung auf einer strukturellen Auswertung mittels Prototypen. Dabei besteht die Möglichkeit, diese Prototypen auf eine Schrift anzupassen, ohne dass ein aufwendiges Training mit vielen Trainingsdaten erfolgen muss.

Die verwendeten Prototypen besitzen stabile Merkmale, die zur Unterscheidung zwischen den Wörtern relevant sind. Dabei ist es nicht erforderlich, dass sämtliche Striche eines Wortes vom Prototypen erfasst werden.

Eine Wort kann nicht als kompakte Anordnung von Strichen betrachtet werden. Es ist aus Buchstaben aufgebaut, deren Abstände zueinander von Wort zu Wort variieren können. Daher wird ein Prototyp aus Abschnitten zusammengesetzt, die horizontal angeordnet sind. Jeder Abschnitt wiederum besteht aus ein bis vier Primitiven, die eine markante Konstellation von Strichen repräsentieren. Während der Erkennung wird versucht, diese Strichanordnung im Kandidaten wieder zu finden.

Es gibt mehrere Möglichkeiten, Prototypen zu erzeugen. Ein kundiger Nutzer kann am besten entscheiden, welche Merkmale für ein Wort in der entsprechenden Schriftform relevant ist und dementsprechend einen Prototypen konstruieren. Diese manuelle Vorgehensweise liefert die besten Ergebnisse, ist allerdings am aufwendigsten.

Dem gegenüber steht die automatische Erzeugung. Dabei müssen vom Nutzer lediglich zwei bis vier typische Vertreter eines Wortes ausgewählt werden. Das System extrahiert aus diesen Beispielen

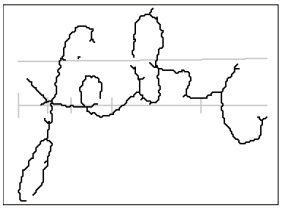
markante Merkmale und erzeugt einen Prototypen. Dabei werden vor allem wichtige Stellen eines Wortes untersucht wie Wortanfang, Ober- und Unterlängen.

Die dritte Möglichkeit, die in dieser Arbeit ansatzweise demonstriert wird, ist die Anpassung existierender Prototypen an Schrift, die sich in bestimmten Parametern geändert hat. Der strukturelle Aufbau der Prototypen ermöglicht gezielte Veränderungen und vermeidet somit ein komplett neues Erzeugen der Prototypen. In dieser Arbeit wurde am Beispiel des Verhältnisses zwischen der Größe von Majuskeln und Minuskeln diese Art der Anpassung durchgeführt.

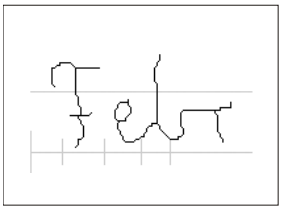
Generische und Adaptierte Prototypen

Aufgrund des Bestrebens, den erforderlichen Rechenaufwand zu minimieren, ist es wünschenswert, die Zahl der Prototypen so gering wie möglich zu halten. Soll ein Prototyp mehrere Schriftstile eines Wortes abdecken, müssen die Merkmale stabil sein und nicht auf individuelle Ausprägungen basieren. Jedoch muss berücksichtigt werden, dass bei größeren Varianzen zwischen den Instanzen eines Wortes die Gemeinsamkeiten kleiner werden. Die Repräsentation durch einen Prototypen wird schwieriger. Es wird notwendig, spezialisierte Prototypen zu erzeugen, um die gewünschte Erkennungsleistung zu erhalten. Nur ein angepasster Prototyp ermöglicht eine stabile Erkennung auch unter schwierigen Bedingungen, wie sie beispielsweise durch Artefakte verursacht werden.

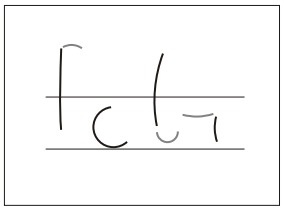
Zur Veranschaulichung sind in Abbildung 3.11 ein generischer und zwei spezialisierte Prototypen dargestellt. Da die Gemeinsamkeiten geringer sind, ist die Zahl der Primitive des generischen Prototypen kleiner und damit auch die Robustheit der Erkennung. Die Prototypen wurden automatisch erzeugt.



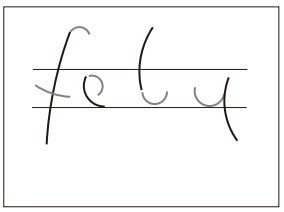
(a) Beispielwort von 1719.



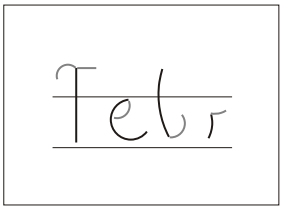
(b) Beispielwort von 1770.



(c) Generischer Prototyp für beide Stile.



(d) Spezialisierter Prototyp für Beispielwort von 1719.



(e) Spezialisierter Prototyp für Beispielwort von 1770.

Abbildung 3.11: Beispiele eines generischen und zwei spezialisierter Prototypen.

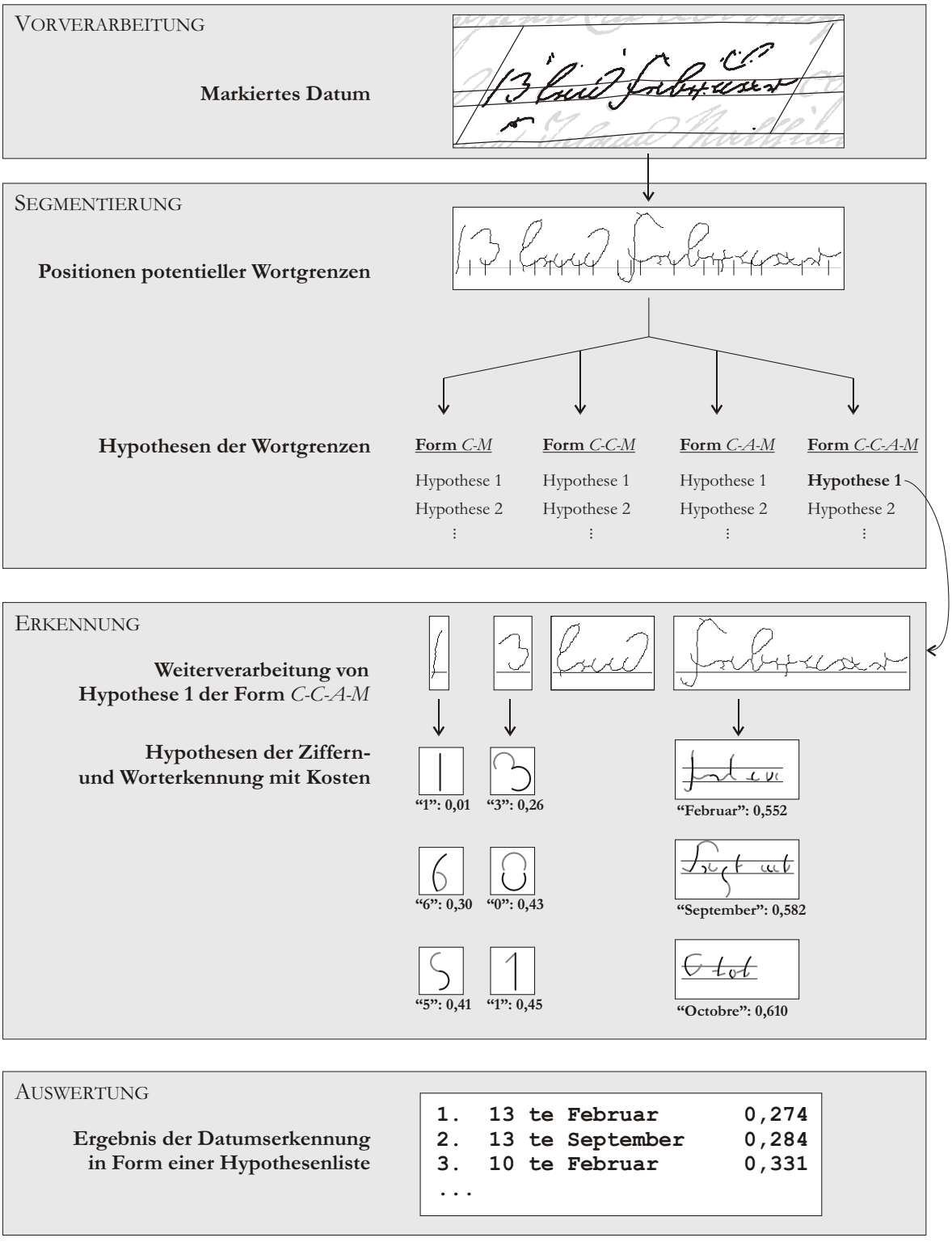


Abbildung 3.12: Ablauf einer Datumserkennung. Nachdem Markieren erfolgt das Bilden der Hypothesen der Wortgrenzen. Die Erkennung wird für die Objekte der wahrscheinlichsten Hypothesen durchgeführt. Abschließend werden die Ergebnisse der Ziffern- und Worterkennung kombiniert. Die Formen der Wortgrenzen (*C-M*,...) werden in Abschnitt 4.1 auf Seite 36 erläutert.

Kapitel 4

Segmentierung des Datums

Mein Freund, die Zeiten der Vergangenheit
sind uns ein Buch mit sieben Siegeln.

Johann Wolfgang von Goethe (1749-1832)

Eine Zeile in ihre Wörter zu zerlegen, ist für Druckschrift aber auch für viele Texte in Schreibschrift aufgrund der eindeutigen Lücken zwischen den Wörtern ohne größere Probleme möglich. Dies gilt jedoch nicht für die Schriften in alten Kirchenbüchern. Wie in Kapitel 2 ausgeführt, können Lücken innerhalb von Wörtern auftreten, die größer sind als zwischen zwei Wörtern oder es kommt zu Wortberührungen. Auf der Basis von geometrischen Analysen lassen sich keine eindeutigen Aussagen über die Positionen der Wortgrenzen machen. Daher wird in diesem Kapitel ein Verfahren vorgestellt, das zusätzliches Wissen über die Struktur des zu erkennenden Teils einbezieht und daraus eine verhältnismäßig kleine Zahl von Hypothesen über die Anordnung der Wörter erstellt [20, 21].

Nachdem die Seite eines Dokuments digitalisiert und entsprechend vorverarbeitet wurde, erfolgt die Segmentierung der Zeilen, wie sie in [17] beschrieben wird. Das Ergebnis dieser Verarbeitung sind die voneinander segmentierten Zeilen mit den entsprechenden Schriftobjekten. Die Einschränkung des in dieser Arbeit gewählten Problems erfordert die Markierung des interessanten Abschnitts – dem Datum. Somit werden dem Modul für die Wortsegmentierung die Schriftobjekte des Datums sowie die Informationen über den Verlauf der Schriftlinien übergeben.

Im Folgenden werden zunächst die Formen erläutert, die ein Eintrag des Datums haben kann. Anschließend wird der Algorithmus detailliert vorgestellt. Dies beinhaltet das Erzeugen der potentiellen Wortgrenzen, das Bewerten der Wortgrenzen durch die Position innerhalb des Datums und durch die Analyse der lokalen Merkmale sowie das Erzeugen der Hypothesen.

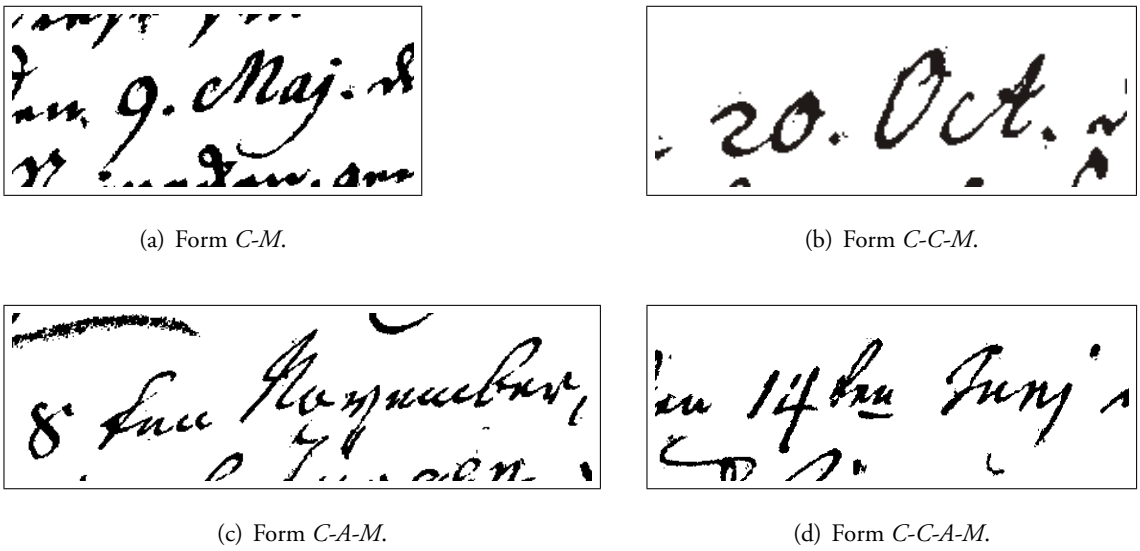
4.1 Formen des Datums

Wie bereits in Abschnitt 2.3 beschrieben, gab es verschiedene Möglichkeiten, das Datum eines Ereignisses anzugeben. In dieser Arbeit wird die Form des Datums behandelt, die sehr häufig in Kirchenbüchern des 18. und 19. Jahrhunderts auftritt. Diese besteht aus dem Tag als arabische Zahl und dem Monat in ausgeschriebener oder abgekürzter Form.

Solche Datumsangaben bestehen aus Elementen, die sich in drei Klassen einordnen lassen: Ziffer (*C*), Artefakt (*A*) und Monatsname (*M*). Unter Artefakte sind Suffixe der Tagesangaben zu verstehen, wie z. B. „te“ oder „ten“. Daraus ergeben sich vier mögliche Formen, wie ein Datum geschrieben werden kann:

- *C-M*
- *C-C-M*
- *C-A-M*
- *C-C-A-M*

Beispiele zu den einzelnen Formen sind in Abbildung 4.1 dargestellt.



(a) Form *C-M*.

(b) Form *C-C-M*.

(c) Form *C-A-M*.

(d) Form *C-C-A-M*.

Abbildung 4.1: Beispiele der vier möglichen Datumsformen.

4.2 Neigungskorrektur

Die Neigungskorrektur ist ein wichtiger Teil der Vorverarbeitung der Wortsegmentierung. Zum Einen ist sie Voraussetzung für eine Suche nach senkrechten Wortgrenzen. Zum Anderen befreit

dieser Schritt alle folgenden Verarbeitungsschritte von einer der vielen Variablen, die eine Handschrift auszeichnen: der Neigung der Schrift.

Es gibt unterschiedliche Verfahren, die Neigung von Handschrift zu bestimmen. So wird beispielsweise bei Vinciarelli et al. [95] und Kavallieratou et al. [34] ein Verfahren vorgestellt, in dem ein Histogramm über die vertikale Dichte der Spalten für mehrere Neigungswinkel errechnet und ausgewertet wird. Da durch die erfolgten Vorverarbeitungsschritte eine Datenstruktur aus Skelettsegmenten vorliegt, wird in unserem Fall eine weniger aufwendige Methode genutzt, die die Lage und Länge dieser Segmente betrachtet, um den Neigungswert zu bestimmen. Sie lieferte für die vorliegenden Daten ausreichend genaue Ergebnisse.

4.2.1 Bestimmen des Neigungswertes

Die Informationen über die Zugehörigkeit der Schriftobjekte zu den einzelnen Zeilen wurden in den vorherigen Schritten gewonnen und dienen dazu, eine Auswahl der Schriftobjekte zu treffen, die mit hoher Sicherheit zu der zu untersuchenden Zeile gehören.

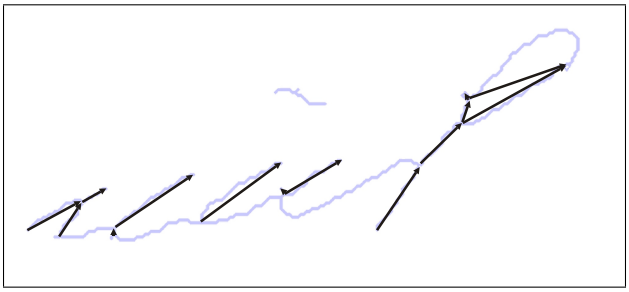


Abbildung 4.2: Vektoren, die zur Bestimmung des Neigungswinkels genutzt werden.

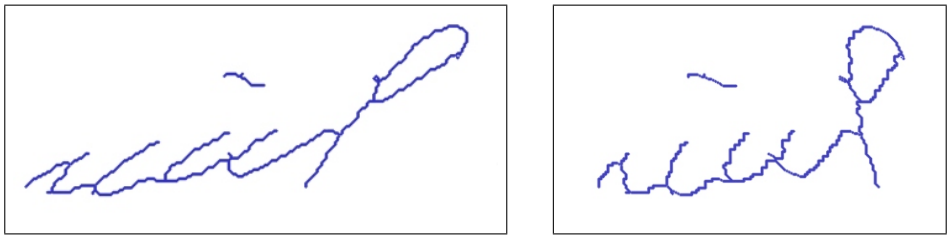
Für alle Skelettsegmente der Schriftobjekte wird der Vektor bestimmt, der sich aus End- und Startpunkt des Segments ergibt. Die Festlegung, dass die Höhe des Startpunktes höchstens so groß ist wie die Höhe des Endpunktes, macht eine Überprüfung des Richtungssinns überflüssig. Es werden nicht alle Vektoren zur weiteren Berechnung betrachtet. Vektoren, die nahezu waagerecht verlaufen, entstehen u. a. durch Ligaturen und können zur Bestimmung der Schriftneigung nichts beitragen. Daher werden nur Vektoren betrachtet, deren Winkel zwischen 20 und 160° liegen. Aus den sich ergebenden n Vektoren mit dem Index $k = 1 \ldots n$ wird aus den Richtungen β_k die durchschnittliche Richtung $\overline{\beta}$ bestimmt, wobei dies in Abhängigkeit der Länge l_k eines Segments erfolgt:

$$\overline{\beta} = \frac{\sum_{k=1}^{n} (\beta_k \cdot l_k)}{\sum_{k=1}^{n} l_k} \tag{4.1}$$

4.2.2 Korrektur der Neigung

Mit Hilfe des Neigungswertes $\overline{\beta}$ wird jeder Punkt px_i des Skeletts in Abhängigkeit seines Abstands von der Basislinie dist (BL, px_i) horizontal um Δx_i verschoben.

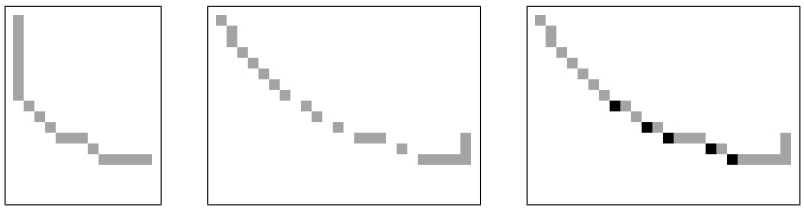
$$\Delta x_i = -\frac{\text{dist}\,(BL, px_i)}{\tan \overline{\beta}} \tag{4.2}$$



(a) Skelett der Schriftobjekte. (b) Skelett nach der Neigungskorrektur.

Abbildung 4.3: Korrektur der Neigung mit Hilfe des zuvor ermittelten Neigungswertes.

Bei Teilen der Segmente, deren Winkel zwischen 90 und 180° liegt (bzw. 270 und 360°), kann es durch die Korrektur zu Unterbrechungen kommen. Diese Stellen werden gesucht und geschlossen, um wieder ein gültiges Skelettsegment zu erhalten (siehe Abbildung 4.4).



(a) (b) (c)

Abbildung 4.4: Die Neigungskorrektur erfordert ein anschließenden Schließen von Lücken innerhalb der Segmente. (a) Segment vor der Korrektur. (b) Segment mit Lücken, die durch die Korrektur entstanden. (c) Segment nach dem Schließen der Lücken.

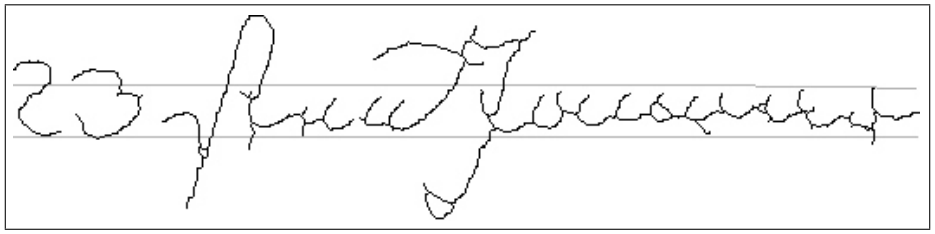
4.3 Potentielle Wortgrenzen

In Abschnitt 2.4.2 wurde erläutert, dass als eine Besonderheit der Schrift in alten Kirchenbüchern die Berührung benachbarter Worte zu nennen ist. So werden in der Regel die Wortgrenzen durch die Existenz größerer Lücken in horizontaler Richtung gekennzeichnet, aber es gibt auch Fälle, in denen die Lücken innerhalb eines Wortes größer sind oder gar keine Lücke zwischen zwei Worten existiert.

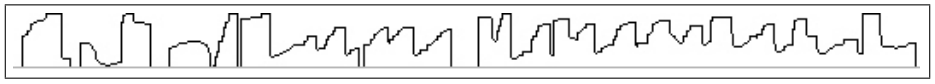
Daher werden zwei Typen von potentiellen Wortgrenzen gebildet:

Potentielle Wortgrenze Typ 1: Bei diesem Typ wird nach horizontalen Lücken gesucht. Dabei wird nur der Zeilenhauptraum betrachtet, um die Probleme eines Teils möglicher Wortberührungen umgehen zu können. So werden alle Wortgrenzen erfasst, die durch Lücken benachbarter Textobjekte gekennzeichnet sind oder bei denen eine Berührung außerhalb des Zeilenhauptraumes existiert.

Potentielle Wortgrenze Typ 2: Wenn sich die Berührung zweier Worte im Zeilenhauptraum befindet, wird die Wortgrenze von diesem zweiten Typ erfasst. Da dies meist durch zu lang gezogene Striche nahe der Basislinie verursacht wird, wird eine potentielle Wortgrenze dieses Typs an jeder Stelle gebildet, an denen die obere vertikale Ausdehnung ein lokales Minimum aufweist und sich im Zeilenhauptraum nur ein Liniensegment befindet.



(a) Skelett der neigungskorrigierten Schriftobjekte.



(b) Oberes Hüllenprofil im Zeilenhauptraums.

(c) Resultierende Positionen potentieller Wortgrenzen.

Abbildung 4.5: Die potentiellen Wortgrenzen werden mit Hilfe des oberen Hüllenprofils gebildet.

Der Raum, in dem sich die Schriftobjekte befinden, wurde durch den Verlauf der Mittellinie und Basislinie in drei Bereiche eingeteilt: den Bereich der Oberlängen, den Bereich der Unterlängen und den Zeilenhauptraum. Für die Suche nach den potentiellen Wortgrenzen werden die Striche der Schriftobjekte betrachtet, die sich im Zeilenhauptraum befinden. Für diese Schriftteile wird die Höhe bestimmt (siehe Abbildung 4.5(b)).

In diesem Höhenprofil wird nach lokalen Minima gesucht. Dabei werden kleinere Höhenschwankungen nicht berücksichtigt. Dies wird erreicht, indem die Tiefe des jeweiligen Tals nach beiden Seiten mindestens $1/5$ des mittleren Abstands $\mathrm{dist}\,(ML, BL)$ der Mittellinie ML von der Basislinie BL betragen muss.

Das Ergebnis dieser Suche beider potentieller Grenztypen ist eine Liste von horizontalen Positionen (siehe Abbildung 4.5(c)). Die Anzahl dieser potentiellen Grenzen liegt in den zur Verfügung stehenden Datumsangaben zwischen 3 und 25.

4.4 Bewertung potentieller Wortgrenzen

Wie schon einleitend in diesem Kapitel beschrieben, stehen durch die Einschränkung des Problems auf die Erkennung des Datums in Kirchenbüchern Informationen zur Verfügung, die als zusätzliches A-priori-Wissen in den Segmentierungs- und Erkennungsprozess einfließen. So werden die potentiellen Wortgrenzen hinsichtlich ihrer *lokalen Eigenschaften* und ihrer *Position* innerhalb des Datums bewertet. Die beiden Wahrscheinlichkeitswerte, die aus diesen Berechnungen hervorgehen, werden gemittelt, um einen Endwert für eine potentielle Grenze zu erhalten. Anschließend werden Hypothesen über die räumliche Aufteilung des Datums in seine Bestandteile aufgestellt.

Aus den vier möglichen Formen, wie ein Datum geschrieben werden kann (siehe Abbschnitt 4.1), ergeben sich acht Arten von Grenzen (B1 ... B8), die zwischen Ziffer (*C*), Artefakt (*A*) und Monatsname (*M*) bestehen:

$$\overset{\mathbf{B1}}{C}|\overset{\mathbf{B2}}{C}|\overset{\mathbf{B3}}{A}|M \quad C|\overset{\mathbf{B4}}{A}|\overset{\mathbf{B5}}{M} \quad C|\overset{\mathbf{B6}}{C}|\overset{\mathbf{B7}}{M} \quad \overset{\mathbf{B8}}{C}|M$$

Da die vorliegende Form des Datums nicht bekannt ist, wird jede potentielle Grenze hinsichtlich jeder der acht Grenzarten untersucht und ein Wahrscheinlichkeitswert ermittelt.

4.4.1 Bewertung lokaler Merkmale

Es stellt sich die Frage, was eine Grenze zwischen zwei Worten bzw. Ziffern von Stellen innerhalb eines Wortes hinsichtlich geometrischer Merkmale unterscheidet. Wie bereits ausgeführt, reicht die Größe einer horizontalen Lücke als einiges Merkmal nicht aus. Daher werden an dieser Stelle vier lokale geometrische Merkmale ($i_1 \ldots i_4$) einer potentiellen Wortgrenze b extrahiert und bewertet (siehe Abbildung 4.6):

Breite der Grenze (i_1). Es wird die linke und rechte Begrenzung ($\mathrm{x_{min}}\,(b)$ und $\mathrm{x_{max}}\,(b)$) des Intervalls und damit die resultierende Breite bestimmt, das die zu untersuchende potentielle Wortgrenze b enthält. Für den Grenztyp I ist $\mathrm{x_{min}}\,(b)$ und $\mathrm{x_{max}}\,(b)$ der Anfang und das Ende der horizontalen Lücke im Zeilenhauptraum. Für den Grenztyp II sind dies die Positionen der nächsten lokalen Maxima des oberen Hüllenverlaufs links bzw. rechts der potentiellen

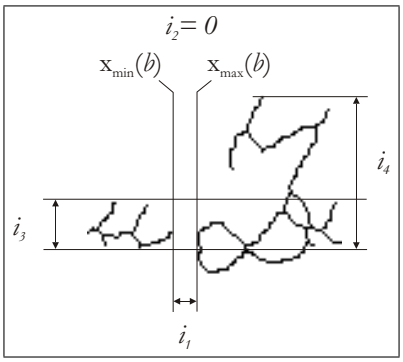


Abbildung 4.6: Die lokalen Merkmale i_1, i_2, i_3 und i_4 einer potentiellen Wortgrenze b werden von einem Multi-Layer Perceptron verarbeitet.

Grenze. Um die Einflüsse der Schriftbreite zu minimieren, wird die ermittelte Breite mittels der Gesamtbreite des Datums w_{date} und der Zahl der gefundenen potentiellen Grenzen N_{pb} normiert. Der Wert i_1 der Breite der Grenze berechnet sich somit folgendermaßen:

$$i_1(b) = (x_{\max}(b) - x_{\min}(b)) \cdot \frac{N_{\text{pb}}}{w_{\text{date}}} \tag{4.3}$$

Zahl der Wortberührungen (i_2). Mit dem Verfahren zur Suche nach potentiellen Wortgrenzen wurde eine Einschränkung getroffen, dass im Zeilenhauptraum höchstens ein Strich durch diese Grenze verlaufen darf. Darüber hinaus ist es jedoch möglich, dass ober- und unterhalb Verbindungen zwischen den Worten bestehen. Je höher diese Zahl ist, umso geringer ist die Wahrscheinlichkeit einer Wortgrenze. Um diesen Zusammenhang zu berücksichtigen, wird die Zahl der Wortberührungen bei der Bewertung erfasst.

Entscheidend ist hierbei, ob es sich um 0, 1 oder 2 Striche handelt. Für höhere Strichzahlen ist keine hohe Signifikanz zu erwarten. Mit anderen Worten: Ob 5 oder 6 Striche eine bestimmte Stelle schneiden, ist für die Frage nach einer potentiellen Wortgrenze irrelevant. Wie in Tabelle 4.1 ersichtlich, wird dementsprechend die Strichzahl auf das Intervall [0,1] übertragen.

Zahl der Striche	Normierter Wert $i_2(b)$
0	0,20
1	0,40
2	0,60
3	0,70
4	0,75
> 4	0,80

Tabelle 4.1: Erzeugen der für das MLP geeigneten Werte aus der Zahl der Striche, die eine horizontale Stelle schneiden.

Schrifthöhe neben der Wortgrenze (i_3, i_4). Die Form der Schrift am Ende oder Anfang eines Wortes kann bei der Suche nach einer Wortgrenze hilfreich sein. So besteht beispielsweise eine hohe Wahrscheinlichkeit, dass ein Wort mit einer Oberlänge beginnt. Daher wird die Höhe der Schrift bestimmt, die sich unmittelbar links ($i_3(b)$) und rechts ($i_4(b)$) der potentiellen Wortgrenze b befindet. Die Größe des Intervalls rechts und links der Grenzposition wird von der Zeichenbreite bestimmt. Diese beträgt maximal das doppelte des durchschnittlichen Abstands der potentiellen Grenzen. Daher ergibt sich die Berechnung der Intervall-Breite w_a wie folgt:

$$w_\mathrm{a} = 2 \cdot \frac{N_\mathrm{pb} + 1}{w_\mathrm{date}} \tag{4.4}$$

Das linke Intervall verläuft somit von $\max(0, \mathrm{x_{min}}(b) - w_\mathrm{a})$ bis $\mathrm{x_{min}}(b)$ und das rechte Intervall von $\mathrm{x_{max}}(b)$ bis $\max(w_\mathrm{date}, \mathrm{x_{max}}(b) + w_\mathrm{a})$.

In jedem dieser beiden Intervalle wird die maximale Höhe der Schriftobjekte bestimmt. Anschließend erfolgt eine Normalisierung, indem durch den Abstand der Mittellinie zur Basislinie dividiert wird.

Für jede potentielle Grenze werden diese vier Eigenschaften ermittelt und mit Hilfe eines neuronalen Netzes bewertet. Dabei handelt es sich um ein Multi-Layer Perceptron (MLP) mit vier Neuronen in der Eingabeschicht, acht Neuronen in der versteckten Schicht und einem Neuron in der Ausgabeschicht. Die Zahl der versteckten Neuronen resultiert aus mehreren Testläufen mit unterschiedlicher Größe der versteckten Schicht. Dabei stand einerseits die Fähigkeit des Netzes im Vordergrund, korrekte Wortgrenzen mit einer minimalen Fehlerrate zu erkennen, während andererseits der Rechenaufwand nur so hoch wie nötig sein sollte.

Das Netz wurde so trainiert, dass ein Wert von 0,2 am Ausgabeneuron erwartet wird, wenn eine falsche Wortgrenze vorliegt und 0,8 wenn es sich um eine korrekte Grenze handelt. Der Grund hierfür liegt in der sigmoiden Transferfunktion der Neuronen. Zu erreichende Ausgabewerte von 0 und 1 sind dafür ungeeignet [45]. Für die weitere Verarbeitung wird der Output o des Netzes auf den Wahrscheinlichkeitswert $\mathrm{p^{NN}}(x_i) \in [0,1]$ normiert:

$$\mathrm{p^{NN}}(x_i) = \begin{cases} 0 & o < 0{,}2 \\ (o - 0{,}2)/0{,}6 & 0{,}2 \leq o \leq 0{,}8 \\ 1 & o > 0{,}8 \end{cases} \tag{4.5}$$

4.4.2 Bewertung der Position

Aus der Breite der einzelnen Bestandteile des Datums ergeben sich typische Positionen für die Grenzen. Mit Hilfe von Testdaten wurden diese Positionen für jede der acht Grenzarten bestimmt. Die

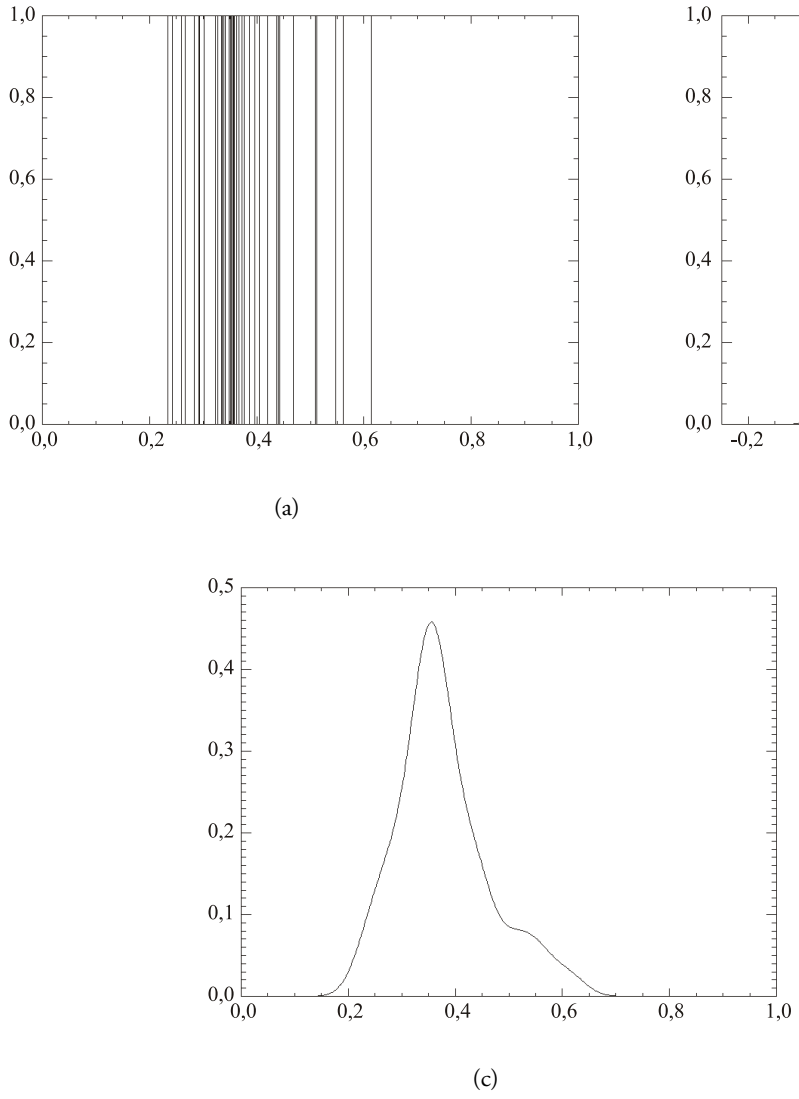


Abbildung 4.7: Berechnung der positionsabhängigen Wahrscheinlichkeitswerte am Beispiel der zweiten Grenze des Grenztyps *C-A-M*. (a) Menge der Trainingsgrenzen $x_{g,j}$. (b) Gaussfunktion mit entprechendem σ_g^2. (c) Resultierender Verlauf der Wahrscheinlichkeitskurve $\mathrm{p}_g^{\mathrm{DC}}(x_i)$.

Position einer Grenze eines Trainingsdatums wird mit der Breite des Datums nomalisiert und liegt damit in dem Intervall [0,1]. Das Resultat ist ein unstetiger Funktionsverlauf, bei dem jede Grenzposition der Testdaten zu einem Peak führt (siehe Abbildung 4.7(a)). Diese Form der Daten lässt sich nicht zur Bewertung der Position einer potentiellen Wortgrenze nutzen. Ein weiterer Verarbeitungsschritt ist erforderlich. Es wird angenommen, dass die Testdaten repräsentative Stichproben darstellen. D. h. jedes Testdatum repräsentiert eine Menge von Daten, deren Grenzpositionen ungefähr mit denen des Testdatums übereinstimmen. Es kann angenommen werden, dass sich diese Grenzpositionen normalverteilt in der Nähe der vorliegenden Grenzpositionen des Testdatums befinden. Dies wird umgesetzt, indem der aus Peaks bestehende Funktionsverlauf mit einer Gausschen Funktion

geglättet wird (siehe Abbildung 4.7(b)). Das Ergebnis ist ein Wahrscheinlichkeitsverlauf über die Breite des Wortes, der zur Bewertung von potentiellen Grenzpositionen genutzt werden kann (siehe Abbildung 4.7(c)). Die Varianz der Gausschen Funktion muss entsprechend angepasst sein und ist zur Zahl der Stichproben umgekehrt proportional (siehe (4.7)). Experimentell wurde hier für den Faktor k ein Wert von 0,04 gefunden. Der Wahrscheinlichkeitswert $\mathrm{p}_g^{\mathrm{DC}}(x_i)$ einer potentiellen Wortgrenze mit der Position $x_i \in [0,1]$ wird für den Grenztyp g folgendermaßen bestimmt:

$$\mathrm{p}_g^{\mathrm{DC}}(x_i) = \frac{1}{N_g} \cdot \sum_{j=1}^{N_g} \exp\left(\frac{(x_i - x_{g,j})^2}{-2\sigma_g^2}\right) \tag{4.6}$$

$$\sigma_g^2 = k \cdot \frac{1}{N_g} \tag{4.7}$$

N_g bezeichnet hierbei die Zahl der Trainingsgrenzen $x_{g,j}$ für den Grenztyp g und $j = 1 \ldots N_g$.

Nachdem die Bewertung der einzelnen potentiellen Grenzen erfolgte, können Hypothesen über die Anordnung mehrerer Wortgrenzen gebildet werden.

4.5 Generieren der Hypothesen

Die Form des zur Erkennung vorliegenden Datums ist unbekannt. D. h. es ist nicht bekannt, ob eine oder zwei Ziffern vorliegen, ob es ein Artefakt gibt oder nicht. Daher werden für jede der vier möglichen Datumsformen alle möglichen Hypothesen aus den N_{pb} potentiellen Grenzpositionen gebildet. Für eine Datumsform mit k potentiellen Grenzen ergeben sich $\binom{N_{\mathrm{pb}}}{k}$ Hypothesen. Für die Datumsform *C-C-A-M* sind dies $\binom{N_{\mathrm{pb}}}{3}$, für Datumsformen *C-C-M* und *C-A-M* $\binom{N_{\mathrm{pb}}}{2}$ und für die Datumsform *C-M* N_{pb} Kombinationen. Der Wahrscheinlichkeitswert einer Hypothese wird aus dem Mittel der Wahrscheinlichkeitswerte der beteiligten Wortgrenzen bestimmt.

Das Ergebnis dieser Berechnungen sind vier Listen mit den Hypothesen über die Positionen möglicher Wortgrenzen. Nachdem diese Listen nach den ermittelten Wahrscheinlichkeiten sortiert wurde, stehen die Hypothesen mit der besten Bewertung am Anfang.

Bisher ist eine Vereinigung der vier Listen auf der Basis der Wahrscheinlichkeiten nicht gelungen. Die Werte der Hypothesen unterschiedlichen Typs lassen sich nicht miteinander vergleichen. Daher werden zunächst die besten Hypothesen der vier Listen zur weiteren Verarbeitung betrachtet.

Kapitel 5

Erkennung von Ziffern

Wer in der Zukunft lesen will,
muss in der Vergangenheit blättern.

André Malraux (1901-1976)

Neben bestimmten Schlüsselwörtern sind es in historischen Aufzeichnungen vor allem die Ziffern, die für Historiker eine hohe Bedeutung besitzen. Seien es nun Lagerbestände, die Dauer einer Haftstrafe, die Größe einer militärischen Einheit oder – wie bei dem hier vorliegenden Problem – das Datum eines Ereignisses. In jedem Fall handelt es sich um präzise Angaben, die mit anderen Quellen abgeglichen werden können. Daher ist es sinnvoll zu untersuchen, inwieweit Ziffern in alten Dokumenten automatisch erkannt werden können. Darüber hinaus stellt sich die Frage, inwieweit Erkenntnisse auch für die Erkennung von Wörtern gewonnen werden können.

In diesem Kapitel wird die Funktionsweise des strukturellen Ansatzes anhand der Erkennung von Ziffern beschrieben. Dieses Verfahren wird in Kapitel 6 zur Erkennung von Wörtern erweitert. Die Vorteile sind die Unabhängigkeit von einem großen Trainingsdatensatz, die Robustheit gegenüber Störungen durch fremde Striche und Lücken im Strichverlauf, die Transparenz der Verarbeitung im Gegensatz zum „Black-Box"-Verhalten von neuronalen Netzen und die Möglichkeit einer Anpassung an eine veränderte Schrift.

In den folgenden Abschnitten werden zunächst die zentralen Begriffe des *Primitivs* und der *approximativen Repräsentation* erläutert bevor anschließend das Verfahren zur Ziffernerkennung detailliert beschrieben wird.

5.1 Primitive

Das hier vorgestellte Verfahren basiert auf dem Erzeugen und Vergleichen von Primitiven, die den Verlauf von Strichen repräsentieren. Es handelt sich hierbei um geometrische Objekte, die durch

wenige Parameter eindeutig beschrieben werden können. In diesem Fall sind dies Kreisbögen, die jeweils einen bestimmten Teil der Striche repräsentieren. Im Gegensatz zur Arbeit von N. A. Khan [38, 37] können diese Kreisbögen einen beliebigen Winkel besitzen, um die Abweichungen der Representation möglichst gering zu gestalten. Kreisbögen werden durch relativ wenige Parameter bestimmt und bieten dadurch eine gute Möglichkeit des Vergleichs. Diese Parameter sind folgende (siehe Abb. 5.1):

- Position des Startpunktes S
- Position des Endpunktes E
- Position des Mittelpunktes M
- Radius[1] r
- Richtungssinn

Darüber hinaus lassen sich für spätere Berechnungen und Bewertungen noch weitere Größen ermitteln:

- Position des Halbpunktes H
- Länge der Strecke $\overline{SE}$
- Länge des Bogens l
- Richtung σ von S zu M
- Richtung ϵ von E zu M
- Winkel des Bogens α

Zwei Sonderfälle bilden zum Einen gerade Strecken ($r = \infty$) und zum Anderen komplette Kreise ($S = E$).

Eine weitere wichtige Eigenschaft eines Primitivs $Prim$ ist dessen Größe. Die Funktion $\mathrm{size}\,(Prim)$ wird für die während der Erkennung stattfindenden Berechnungen genutzt, um den Einfluss der Primitivgröße zu verhindern. Sie repräsentiert die größte Ausdehnung der Bounding Box, Breite $\mathrm{w}\,(Prim)$ oder Höhe $\mathrm{h}\,(Prim)$ (siehe Formel 5.1). Es ist ein ausreichend genauer und schnell berechneter Wert, der die Größe des Primitivs repräsentiert.

$$\mathrm{size}\,(Prim) = \max\left(\mathrm{w}\,(Prim), \mathrm{h}\,(Prim)\right) \tag{5.1}$$

[1] Aus der Position der drei Punkte kann der Radius bestimmt werden. Er gehört aber dennoch zu den grundlegenden Eigenschaften.

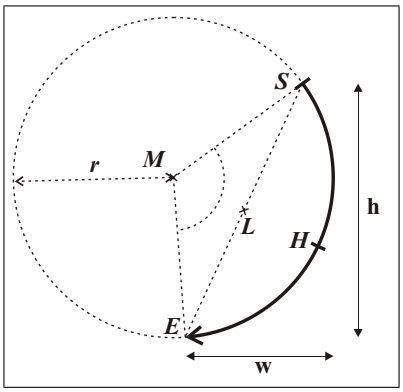


Abbildung 5.1: Parameter und Merkmale eines Kreisbogens.

5.2 Approximative Repräsentation

Die zur Erkennung genutzten Prototypen bestehen aus Primitiven, die Striche des jeweiligen Objektes repräsentieren. Während der Erkennung wird bewertet, ob und inwieweit die Strichkonstellationen des Prototypen im Kandidaten wiedergefunden werden können. Dazu ist es erforderlich, die Striche des Kandidaten durch Primitive approximativ nachzubilden. Dies basiert auf den Segmenten des zuvor erzeugten Skeletts der Schriftobjekte. Die Generierung der Primitive erfolgt in zwei Schritten. Zuerst wird der Verlauf der Skelett-Segmente durch viele kurze Primitive nachgezeichnet. Kann der Verlauf zweier Primitive kombiniert werden, wird dieser durch ein bis zwei neue Primitive repräsentiert. Dies geschieht dann, wenn die Abweichung vom Verlauf der Striche innerhalb einer festgelegten Toleranz liegt. Der zweite Schritt beinhaltet mehrere Durchläufe dieses Fusionierens von Primitiven. Der Prozess bricht ab, sobald kein weiteres Primitiv auf diese Weise gebildet werden kann.

5.2.1 Erste Approximation

In Abhängigkeit von der Auflösung während des Digitalisierens, wird der Verlauf der Striche durch eine Vielzahl von Skelettpunkten repräsentiert. Es wäre möglich, die Primitive der ersten Approximation durch jeweils drei benachbarte Skelettpunkte zu erzeugen. Dies würde zu einer hohen Zahl resultierender Primitive führen. Der Rechenaufwand der folgenden Schritte würde stark ansteigen. Diese geringe Größe der Primitive ist nicht notwendig, da ein auf das Papier gebrachter Federzug in Abhängigkeit seiner Dicke nur einen bestimmten minimalen Kurvenradius besitzen kann. Daher wird die Länge der Primitive der ersten Approximation in Abhängigkeit der jeweiligen Strichdicke festgelegt. Steht die Information der Dicke für jeden Strich zur Verfügung, kann sie genutzt werden; ansonsten wird ein konstanter Wert festgelegt, der der vorliegenden Schrift entspricht. Für ein Primitiv der ersten Approximation gilt folgende Maximallänge $l_{\max}^{\mathrm{EA}}$ in Abhängigkeit der Strichdicke:

$$l_{\max}^{\mathrm{EA}} = 2 \cdot Strichdicke \tag{5.2}$$

Selbst für recht kleine Schriften mit einer Minuskelhöhe von 1,4 mm liefert der Faktor 2 eine ausreichend genaue Approximation.

Zur Reduzierung des Rechenaufwands kann eine Mindestlänge eingeführt werden, damit auch bei sehr dünnen Linien die Zahl der Primitive nicht höher als nötig wird. In der aktuellen Umsetzung wird hier ein Wert von $\mathrm{dist}\,(ML, BL)\,/4$ verwendet.

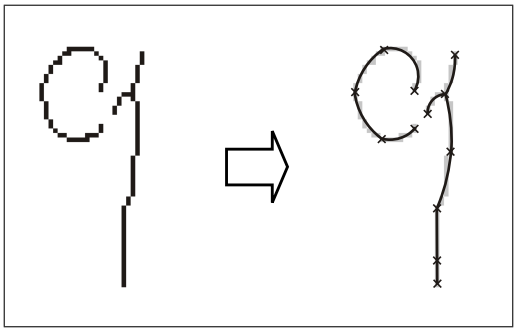


Abbildung 5.2: Erzeugen der ersten Approximation. In diesem Beispiel werden aus den vier Skelettsegmenten zehn Primitive erzeugt.

Die Verarbeitung erfolgt segmentweise. Ein Skelettsegment mit der Länge s ist eine Anordnung von s Skelett-Pixeln px_i mit $i = 0 \ldots s-1$. Ist s nicht größer als $l_{\max}^{\mathrm{EA}}$ wird nur ein Primitiv zur Approximation dieses Segments erzeugt. Dabei bestimmen Anfangs- und Endpunkt des Segmentes (px_0, px_{s-1}) den Anfangs- und den Endpunkt des Primitivs (S, E). Das mittlere Skelett-Pixel $px_{\lceil s/2 \rceil}$ bestimmt die Position des Halbpunktes H.

Bei längeren Segmenten wird deren Verlauf durch mehrere hintereinander angeordnete Primitive approximiert. Dabei ist der Endpunkt eines Primitivs jeweils der Anfangspunkt des nächsten. Der Endpunkt dieses Primitivs wird durch $l_{\max}^{\mathrm{EA}}$ oder px_{s-1} bestimmt. Das Resultat dieses ersten Schrittes ist in Abbildung 5.2 in einem Beispiel dargestellt.

5.2.2 Fusionieren von Primitiven

Nachdem das Bilden der ersten Approximation abgeschlossen ist, beginnt die Phase des Fusionierens von Primitiven. Eine *Fusion* zweier Primitive bedeutet die vollständige oder teilweise approximative Representation durch ein neues, meist größeres Primitiv. Ist die Abweichung des neuen Primitivs zum ursprünglichen Verlauf der Striche zu groß, wird es nicht gebildet. Dies ist vor allem dann der Fall, wenn der Abstand der beiden Quell-Primitive zu groß ist, sodass in vielen Fällen ein Versuch einer Fusion nicht durchgeführt werden muss.

Zur Bewertung des Abstands wird die Minuskelhöhe h_{Min} der vorliegenden Schriftgröße herangezogen. Ein Punkt liegt nahe eines Primitivs, wenn der Abstand kleiner ist, als der Grenzwert th_{P}.

$$th_{\mathrm{P}} = h_{\mathrm{Min}}/5 \tag{5.3}$$

Diese Einschränkung dient in erster Linie der Reduzierung des Rechenaufwands. Die hier vorgeschlagene Bestimmung von th_{P} lieferte eine gute Approximation für alle Daten. Ist der Grenzwert zu groß, steigt die Zahl der Versuche einer Fusion und damit auch die Rechenzeit unnötig an. Wird th_{P} zu klein gewählt, werden Lücken im Strichverlauf nicht überbrückt.

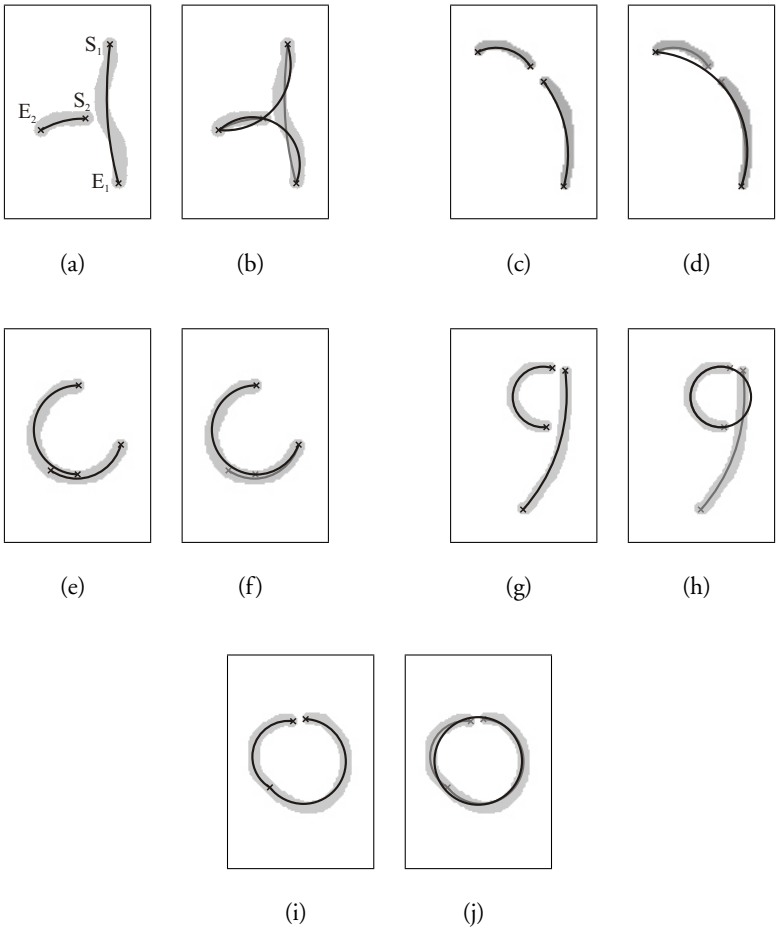
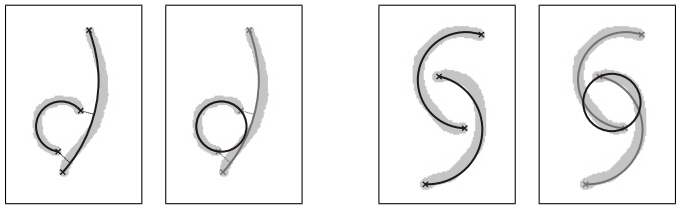


Abbildung 5.3: Unterschiedliche Konstellationen benachbarter Primitive (a), (c), (e), (g), (i) müssen unterschiedlich verarbeitet werden, damit neue Primitive durch Fusionieren gebildet werden können (b), (d), (f), (h), (j).

Betrachtet werden die Punkte S und E zweier Primitive $Prim_1$ und $Prim_2$. Befindet sich mindestens einer der vier Punkte (S_1, E_1, S_2, E_2) in der Nähe des jeweils anderen Primitivs, wird eine Fusion versucht. Abhängig von der Konstellation der beiden Kreisbögen ergeben sich unterschiedliche Formen der Fusion, die ein oder zwei neue Primitive hervorbringt. Welche Fusion stattfindet,



(a) (b) (c) (d)

Abbildung 5.4: Konstellationen, die nicht auftreten, da das neue Primitiv bereits in einem vorherigen Schritt erzeugt wurde.

hängt davon ab, welche Start-/Endpunkte nach der oben genannten Definition dem jeweils anderen Kreisbogen nahe liegen. Im Folgenden werden die unterschiedlichen Situationen geschildert, wenn sich ein bis vier Start-/Endpunkte der Primitive in der Nähe des jeweils anderen Primitivs befinden – der Abstand also kleiner ist als th_{P}.

Ein Start-/Endpunkt befindet sich in der Nähe des anderen Primitivs. Dieser Fall ist in Abbildung 5.3(a) dargestellt. Mit der Position der vier Start- und Endpunkte werden zwei neue Kreisbögen erzeugt. Durch drei Punkte wird ein Kreisbogen eindeutig definiert. In diesem Fall sind dies die Punkte S_1, S_2, E_2 und E_1, S_2, E_2. Das Ergebnis dieser Fusion ist in Abbildung 5.3(b) zu sehen.

Zwei Start-/Endpunkte befinden sich in der Nähe des anderen Primitivs. Von insgesamt vier theoretisch möglichen Konstellationen werden nur zwei betrachtet, da die in Abbildung 5.4(b) und 5.4(d) dargestellten Resultat-Primitive schon in vorhergehenden Durchläufen erzeugt wurden. Wenn sich zwei Start-/Endpunkte nahe sind, wird ein Kreisbogen erzeugt, der den Verlauf beider Primitive komplett nachbildet (Abbildungen 5.3(c), 5.3(d)). Die Parameter des neuen Primitivs werden von den beiden äußeren Start-/Endpunkten bestimmt, sowie der Position des neuen Halbpunktes auf einem der beiden Quell-Primitive. Ebenso wird verfahren, wenn eine Überlappung wie in Abbildung 5.3(e) vorliegt. Der für die Definition eines Kreises notwendige dritte Punkt wird dabei aus den innen liegenden Start-/Endpunkten ausgewählt. Ist ein Primitiv mindestens doppelt so lang wie das andere, so wird dessen Halbpunkt gewählt.

Drei Start-/Endpunkte befinden sich in der Nähe des anderen Primitivs. Ein Kreis wird erzeugt (Abbildungen 5.3(g), 5.3(h)). Mit Hilfe des Primitivs, dessen Start- und Endpunkt in der Nähe des anderen Primitivs liegt, werden Position und Radius des Kreises festgelegt.

Vier Start-/Endpunkte befinden sich in der Nähe des anderen Primitivs. Ein Kreis wird aus dem Verlauf der beiden Primitive gebildet (Abbildungen 5.3(i), 5.3(j)).

In einem Durchlauf werden die paarweisen Kombinationen aller neu hinzugekommenen Primitive im Hinblick auf eine mögliche Fusion getestet. Zu Beginn sind dies die Primitive der ersten

Approximation. Eine Fusion ist erfolgreich, wenn das resultierende Primitiv nicht weiter vom ursprünglichen Verlauf der Striche abweicht als der Grenzwert th_{P}. Die resultierenden Primitive sind der Gegenstand des nächsten Durchlaufs. Konnten keine neuen Primitive erzeugt werden, bricht das Verfahren ab.

Die Ergebnisse der einzelnen Durchläufe werden in einer Liste zusammengefasst, sodass eine Liste der ersten Repräsentation und eine zweite Liste der durch Fusion entstandenen Primitive vorliegen. Für die weitere Verarbeitung ist es von Bedeutung, dass diese Primitive Verweise auf die Primitive der ersten Approximation besitzen. Dadurch ist es möglich, schnell die Länge eines Primitivs abzuschätzen oder gemeinsame Teile zweier Primitive festzustellen.

5.3 Prototypen

Die Merkmale eines zu erkennenden Objektes werden in Prototypen gespeichert. Merkmale sind in diesem Falle die räumliche Anordnung von Strichen. Diese werden durch ein bis vier Primitive approximativ repräsentiert. Grundsätzlich sollten vor allem markante und stabile Striche als Merkmale erfasst werden. Bei relativ kleinen Objekten wie Ziffern ist es erforderlich, alle Striche des Objektes durch den Prototypen zu repräsentieren. Desweiteren ist es aufgrund der geringen Zahl von Merkmalen erforderlich, die Objekte möglichst genau nachzubilden. Daher wurden pro Ziffer jeweils zwei bis fünf Prototypen erzeugt.

Der Prototyp einer Ziffer besteht aus einem bis vier Primitiven. Eines dieser Primitive übernimmt die Rolle des so genannten *Hauptprimitivs*. Über dieses Primitiv erfolgt die Anpassung der Position und der Größe zwischen Prototyp und Kandidat. Alle weiteren Primitive werden in Relation zu diesem Hauptprimitiv bewertet. Im Folgenden werden Hauptprimitive mit dem Index 1 versehen.

Die hier verwendeten Prototypen wurden manuell erstellt. D. h. es wurden Primitive per Hand angeordnet, um die Striche der Ziffern nachzubilden. In anschließenden Tests wurden die bestehenden Prototypen korrigiert und verfeinert, indem bei Fehlleistungen die Ursache der nicht korrekten Erkennung ermittelt wurde – ein Vorzug des strukturellen Verfahrens. Es ist ebenfalls möglich, die Bildung der Prototypen zu automatisieren, wie dies für die Erkennung von Worten in Kapitel 6 beschrieben wird.

Steht ein Satz Prototypen zur Verfügung, können mittels Matching Kandidaten erkannt werden.

5.4 Matching

Nachdem für einen Kandidaten die approximative Repräsentation erzeugt wurde, kann zwischen Prototypen und Kandidaten nach Ähnlichkeiten gesucht werden.

Eine *Korrespondenz* zwischen einem Prototypen P mit k Primitiven $Prim_i^{\mathrm{P}}$ mit $i = 1 \ldots k$ und dem Kandidaten C mit n Primitiven $Prim_j^{\mathrm{C}}$ mit $j = 1 \ldots n$ ist eine Funktion $\mathrm{m}(i) = j$, die jedem Primitiv $Prim_i^{\mathrm{P}}$ des Prototypen ein Primitiv $Prim_j^{\mathrm{C}}$ des Kandidaten zuweist. Die Abweichung zwischen zwei korrespondierenden Primitiven wird durch Kosten repräsentiert, die eine Aussage über die Ähnlichkeit zwischen Prototyp und Kandidat ermöglichen.

Während des Matchings wird nach der besten Korrespondenz zwischen dem Prototypen und dem Kandidaten gesucht. Dies ist die Korrespondenz mit den geringsten Kosten. Anschließend werden die Kosten zwischen den besten Korrespondenzen der unterschiedlichen Prototypen verglichen. Daraus entsteht eine Reihenfolge von Prototypen, die eine Aussage über die Wahrscheinlichkeiten der einzelnen Ziffern für den Kandidaten zulässt.

Die dazu notwendige Kostenberechnung muss so gestaltet sein, dass die Ergebnisse der einzelnen Prototypen vergleichbar sind.

5.5 Kostenberechnung

Während des Matchings werden Korrespondenzen gebildet und bewertet. Die Kosten c^{cipher} für eine Hypothese einer Ziffer setzen sich aus drei Teilkosten zusammen, die durch Faktoren angepasst und addiert werden. Im Wesentlichen werden die Kosten durch das Matching der Primitive des Prototypen auf eine Untermenge der Primitive des Kandidaten bestimmt. Aber auch die Approximationskosten der Primitive des Kandidaten sowie ungenutzte Striche fließen in die Bewertung ein:

Kosten für die Korrespondenz $c^{\mathrm{m}}\left(Prim^{\mathrm{P}}, Prim^{\mathrm{C}}\right)$ erfasst die Abweichungen zwischen den beiden Primitiven $Prim^{\mathrm{P}}$ und $Prim^{\mathrm{C}}$ im Bezug auf Translation, Rotation, Skalierung und Verformung bzw. Differenz der Kreisteile.

Kosten für die Approximation $c^{\mathrm{ap}}\left(Prim^{\mathrm{P}}, Prim^{\mathrm{C}}\right)$ entstehen aus der Abweichung des Primitivs $Prim^{\mathrm{C}}$ vom ursprünglichen Verlauf des Strichs.

Kosten ungenutzter Teile c^{us} entstehen durch Striche des Kandidaten, die für die Korrespondenz nicht genutzt werden.

Die Kosten für ungenutzte Teile c^{us} resultieren aus der Tatsache, dass ein Prototyp nur Aussagen über Striche enthält, die an bestimmten Positionen auftreten. Es gibt jedoch keine Informationen über Striche, die nicht auftreten sollten. Daher ist eine Bewertung aller nicht genutzten Striche des Kandidaten erforderlich.

Die Gesamtkosten eines Matchings zwischen einem Prototypen und einem Kandidaten ergeben sich aus den in Abbildung 5.5 aufgeführten Teilkosten. In welchem Verhältnis diese Einfluss auf das Gesamtergebnis nehmen, wird durch insgesamt sieben Faktoren (f^{ap}, f^{tr}, f^{sc}, f^{ro}, f^{sh}, f^{wp},

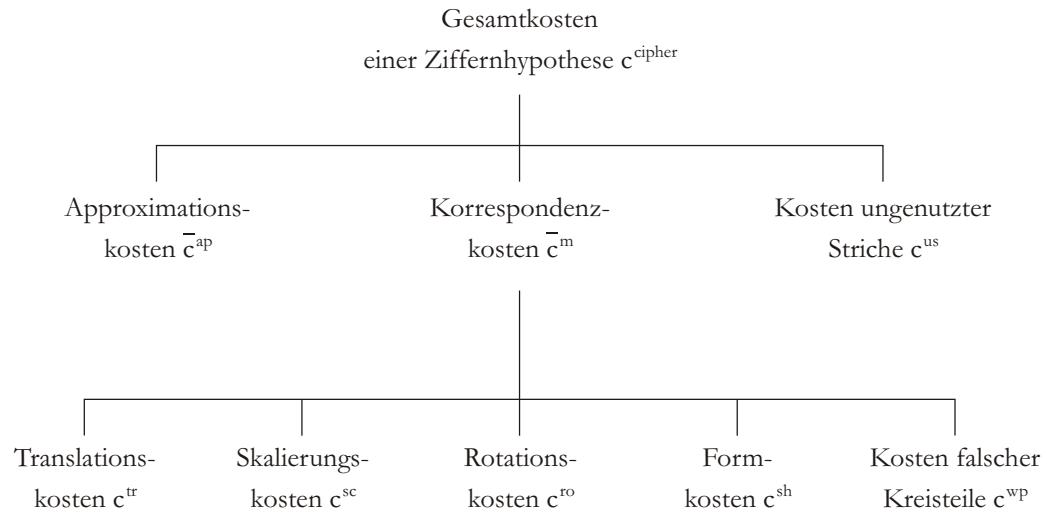


Abbildung 5.5: Übersicht über die Zusammensetzung der Kosten einer Korrespondenz zwischen Primitiven des Prototypen und Primitiven des Kandidaten.

$f_{\mathrm{z}}^{\mathrm{us}}$) geregelt. Die durch Tests ermittelten Werte dieser Faktoren sind in Abschnitt 7.3.1 auf den Seiten 83 ff. detailliert dargestellt.

Für stärker gestörte Schriften, bei denen es häufiger zu Artefakten kommt, die durch benachbarte Zeilen verursacht werden, ist es hilfreich, die Position des Objekts relativ zu den rekonstruierten Textlinien zu bewerten. Die Funktion $\mathrm{pos}\,(\mathrm{m}, BL, ML)$ überprüft mehrere Bedingungen über Größe und Position des Objekts im Bezug auf die Basislinie BL und Mittellinie ML und führt gegebenenfalls zu einer Erhöhung der Kosten (siehe Abschnitt 5.5.4).

Für eine bestimmte Korrespondenz m zwischen einem Prototypen mit k Primitiven einerseits und einem Kandidaten andererseits errechnen sich die Gesamtkosten c^{cipher} folgendermaßen:

$$\overline{c^{\mathrm{ap}}} = \frac{\sum_{i=1}^{k} c^{\mathrm{ap}}\left(Prim_{m(i)}^{\mathrm{C}}\right) \cdot l_i^P}{\sum_{i=1}^{k} l_i^P} \tag{5.4}$$

$$\overline{c^{\mathrm{m}}} = \frac{\sum_{i=1}^{k} c^{\mathrm{m}}\left(Prim_i^{\mathrm{P}}, Prim_{m(i)}^{\mathrm{C}}\right) \cdot l_i^P}{\sum_{i=1}^{k} l_i^P} \tag{5.5}$$

$$c^{\mathrm{cipher}} = (1 - f_{\mathrm{z}}^{\mathrm{us}})\,(\overline{c^{\mathrm{ap}}} + \overline{c^{\mathrm{m}}}) \cdot \mathrm{pos}\,(\mathrm{m}, BL, ML) + f_{\mathrm{z}}^{\mathrm{us}} c^{\mathrm{us}} \tag{5.6}$$

Die Normierung über die Längen l^P aller Prototyp-Primitive ist erforderlich, damit der Einfluss eines Primitivs auf die Kosten in Relation zu dessen Größe steht. Desweiteren wird dadurch die Vergleichbarkeit zwischen den Prototypen gewährleistet.

Der Wertebereich der Gesamtkosten c^{cipher} ist das Intervall [0,1]. Die Summe aus $\overline{c^{\text{ap}}}$ und $\overline{c^{\text{m}}}$ vereint die akkumulierten Teilkosten, die durch unterschiedliche Abweichungen entstehen. Sobald diese Summe die Grenze von 1,0 erreicht hat, ist die Unähnlichkeit zu groß. Ein weiteres Betrachten dieser Korrespondenz ist unnötig. Tritt an einer Stelle der Berechnungen ein Teilkostenwert größer gleich 1,0 auf, wird die Kostenberechnung für diese Korrespondenz abgebrochen. Diese maximale Unähnlichkeit wird mit einer nicht gefundenen Korrespondenz gleichgesetzt.

Es erfolgt während der Teilkostenberechnung somit eine Deckelung des Kostenwertes auf 1,0. Für die Berechnung der Gesamtkosten c^{cipher} ist ein Addieren der Kosten der nicht genutzten Striche c^{us} nicht möglich, da sonst in Situationen mit einem hohen Anteil an fremden Strichen gültige Lösungen mit Kosten größer 1 verworfen würden. Daher wird der Faktor f_{z}^{us}, dessen Wertebereich im Intervall [0,1] liegt, so verwendet, wie es in Gleichung 5.6 zu sehen ist[2]. Er definiert das Verhältnis zwischen den Kosten des Matchings und den Kosten der ungenutzten Striche.

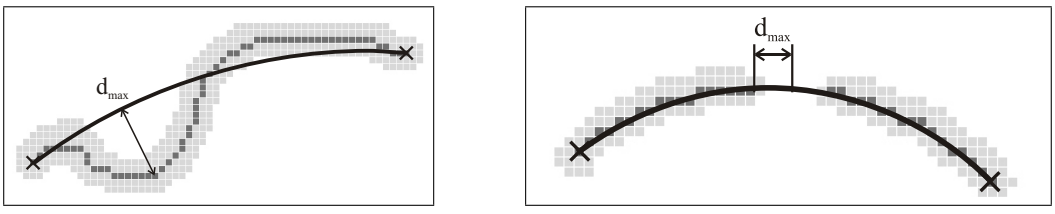
5.5.1 Kosten für die Approximation

Ein Primitiv $Prim^{\text{C}}$ bildet den Verlauf eines oder mehrerer Striche des Kandidaten nach. Gebildet wird es auf der Grundlage des zuvor erzeugten Skeletts. Die größte auftretende Abweichung $d_{\max}\left(Prim^{\text{C}}\right)$ zwischen Strich und Primitiv wird bestimmt und in das Verhältnis zur Primitivgröße $\text{size}\left(Prim^{\text{C}}\right)$ des Primitivs gesetzt. Die Gleichung für die Approximationskosten c^{ap} lautet somit wie folgt:

$$\text{c}^{\text{ap}}\left(Prim^{\text{C}}\right) = \left(\frac{d_{\max}\left(Prim^{\text{C}}\right)}{\text{size}\left(Prim^{\text{C}}\right)}\right)^2 \cdot f^{\text{ap}} \tag{5.7}$$

Zwei maximale Abstände können zwischen Primitiv und Strich bestimmt werden: *(1)* Der vom Primitiv am weitesten entfernte Skelettpunkt. *(2)* Der Punkt auf dem Primitiv-Bogen, der am wei-

[2] Das „z“ bei der Bezeichnung des Faktors f_{z}^{us} dient zur Unterscheidung zum entsprechenden Faktor f_{w}^{us} der Worterkennung. Wie später noch erläutert wird, ist dieser Faktor als einziger der vorliegenden Erkennungsaufgabe anzupassen.



(a) (b)

Abbildung 5.6: Maximale Abweichung zwischen Skelett und Primitive. (a) Zwischen einem Skelettpunkt und dem Primitiv. (b) Zwischen einem Punkt auf dem Primitiv-Bogen und dem Skelett.

testen vom approximierten Skelett entfernt ist. Zur Verdeutlichung sind die beiden Abstände in Abbildung 5.6 dargestellt. Das Maximum dieser beiden Abstandswerte wird durch $d_{\max}\left(Prim^{\mathrm{C}}\right)$ repräsentiert.

5.5.2 Kosten zwischen korrespondierenden Primitiven

Nachdem eine Korrespondenz zwischen dem Prototypen und dem Kandidaten gefunden wurde, gibt es zu jedem Prototyp-Primitiv $Prim^{\mathrm{P}}$ ein Kandidaten-Primitiv $Prim^{\mathrm{C}}$. Zwischen jedem Primitivpaar werden die Primitivkosten berechnet, die durch Abweichungen in Position, Größe, Ausrichtung und Form entstehen.

Auf der Grundlage der beiden Hauptprimitive $Prim_1^{\mathrm{P}}$ und $Prim_1^{\mathrm{C}}$ wird eine Skalierung und Translation bestimmt, sodass zwischen diesen beiden Primitiven keine Abweichungen dieser Art auftreten. D. h. Größe und Position aller weiteren Primitive der Korrespondenz werden vor der Kostenberechnung entsprechend korrigiert. Der für die Translation genutzte *Bezugspunkt* ist jeweils einer der drei Punkte des Hauptprimitivs: S, E oder M. Welcher Punkt gewählt wird, hängt von der Form und Position des Hauptprimitivs ab.

Die Form eines Primitivs variiert innerhalb zweier Extrema: der geraden Linie und dem Kreis. Solche unterschiedlichen geometrischen Objekte werden durch unterschiedliche Arten von Parametern beschrieben. Daher wird die Art und Weise, wie zwei Primitive miteinander verglichen werden, auf zwei Wegen durchgeführt:

Berechnungsweg I : Die Primitive werden als Strecken mit Anfangs- und Endpunkt betrachtet, die gegebenenfalls leicht gekrümmt sind. Die Punkte S, H und E werden zur Bestimmung der Abweichung herangezogen.

Berechnungsweg II: Die Primitive werden als Kreise oder Kreisausschnitte betrachtet. Die Position des Kreismittelpunktes M, der Radius r sowie die Richtungen σ und ϵ werden verglichen.

So wie der Mensch beim Bestimmen von Ähnlichkeiten unbewusst unterschiedliche Strategien testet, werden auch bei der Bestimmung der Ähnlichkeit zwischen zwei Primitiven diese beiden Wege der Kostenbestimmung durchgeführt. Ist ein Weg des Vergleichens für die vorliegenden Primitive ungünstig, so entstehen zu hohe Kosten, die den Grad der Ähnlichkeit nicht repräsentieren. Daher werden die Werte beider Wege ermittelt und das Ergebnis mit den kleineren Kosten ausgewählt. Bei Primitiven mit einem kleineren Kreisausschnitt ist dies Weg I, bei größeren Kreisausschnitten ist es Weg II.

Sollte sich im Folgenden die Berechnung von Teilkosten zwischen Weg I und Weg II unterscheiden, so wird dies durch die Indizes „I“ und „II“ verdeutlicht.

Folgenden Teilkosten werden zur Berechnung der Abweichung zwischen zwei Primitiven bestimmt:

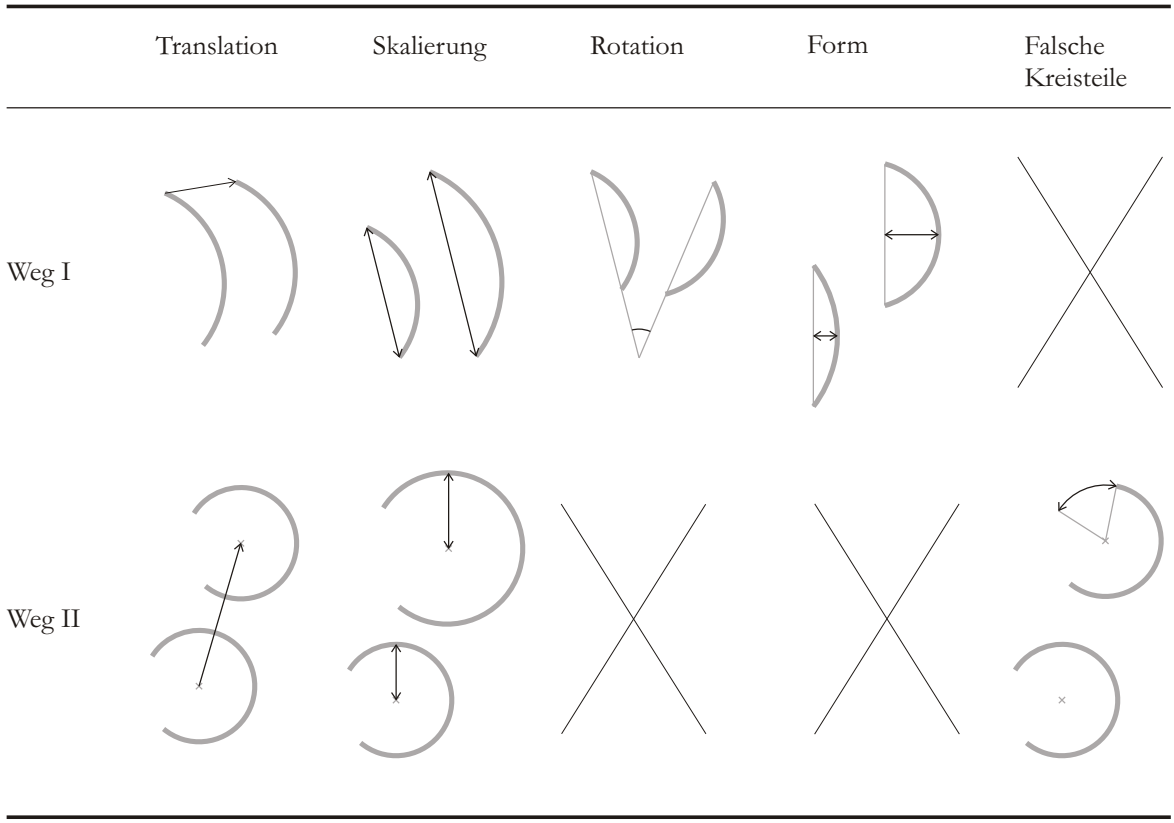


Abbildung 5.7: Die Berechnung der erforderlichen Transformationen zwischen zwei Primitiven erfolgt erfolgt auf zwei Wegen. In Abhängigkeit der vorliegenden Krümmung zweier Primitive liefert der eine oder der andere Weg ein besseres Ergebnis.

- Räumliche Distanz, $c^{\mathrm{tr}}\left(Prim^{\mathrm{P}}, Prim^{\mathrm{C}}\right)$, *Translation*
- Größenunterschied, $c^{\mathrm{sc}}\left(Prim^{\mathrm{P}}, Prim^{\mathrm{C}}\right)$, *Skalierung*
- Richtungsunterschied, $c^{\mathrm{ro}}\left(Prim^{\mathrm{P}}, Prim^{\mathrm{C}}\right)$, *Rotation*
- Krümmungsunterschied, $c^{\mathrm{sh}}\left(Prim^{\mathrm{P}}, Prim^{\mathrm{C}}\right)$, *Form*
- Unterschiedliche Kreisteile, $c^{\mathrm{wp}}\left(Prim^{\mathrm{P}}, Prim^{\mathrm{C}}\right)$, *Falsche Teile*

Abbildung 5.7 verdeutlicht die Formen der möglichen Transformationen zwischen zwei Primitiven. Es ergibt sich folgende Addition der Teilkosten:

$$c_{\mathrm{I}}^{\mathrm{m}} = c_{\mathrm{I}}^{\mathrm{tr}} + c^{\mathrm{sc}} + c_{\mathrm{I}}^{\mathrm{ro}} + c_{\mathrm{I}}^{\mathrm{sh}} \tag{5.8}$$

$$c_{\mathrm{II}}^{\mathrm{m}} = c_{\mathrm{II}}^{\mathrm{tr}} + c^{\mathrm{sc}} + c_{\mathrm{II}}^{\mathrm{wp}} \tag{5.9}$$

$$c^{\mathrm{m}} = \min\left(c_{\mathrm{I}}^{\mathrm{m}}, c_{\mathrm{II}}^{\mathrm{m}}\right) \tag{5.10}$$

Im Folgenden werden die einzelnen Teilkosten zwischen zwei Primitiven genauer erläutert. Bevor die Berechnung beginnt, wird die richtige Zuordnung zwischen den Start- und Endpunkten der Primitive bestimmt. Dazu wird die Winkeldifferenz der Strecken $\overline{S^{\mathrm{P}}E^{\mathrm{P}}}$ und $\overline{S^{\mathrm{C}}E^{\mathrm{C}}}$ berechnet. Sie muss kleiner sein als $\pi/2$ anderenfalls werden die Bezeichnungen der Punkte S^{C} und E^{C} getauscht.

Die Teilkosten der Translation und der Skalierung sind von der Primitivgröße abhängig. Daher werden diese entsprechend normiert. Der Normierungswert nv wird von der größten Ausdehnung des kleineren Primitivs unter Nutzung der size ()-Funktion (siehe Seite 46) bestimmt:

$$nv = \min\left(\mathrm{size}\left(Prim^{\mathrm{P}}\right), \mathrm{size}\left(Prim^{\mathrm{C}}\right)\right) \tag{5.11}$$

Die Kosten der Positionsabweichung erfassen die erforderliche Translation. Als Translationspunkt[3] werden auf dem ersten Berechnungsweg die Start- und Endpunkte der Primitive in Betracht gezogen. Ausgewählt wird der dem Bezugspunkt *deutlich* näher liegende Punkt. D. h. der Distanzunterschied muss größer sein als die halbe Primitivlänge des Prototypen $l^{\mathrm{P}}/2$, ansonsten wird die kleinere Translation gewählt. Der zweite Weg wählt den Kreismittelpunkt als Translationspunkt.

Zur besseren Lesbarkeit der Gleichung 5.14 sei s die Distanz zwischen den beiden Startpunkten und e die Distanz zwischen den Endpunkten der beiden Primitive:

$$s = \left|\overline{S^{\mathrm{P}}S^{\mathrm{C}}}\right| \tag{5.12}$$

$$e = \left|\overline{E^{\mathrm{P}}E^{\mathrm{C}}}\right| \tag{5.13}$$

$$\mathrm{c}_I^{\mathrm{tr}}\left(Prim^{\mathrm{P}}, Prim^{\mathrm{C}}\right) = \begin{cases} \left(\frac{s}{nv}\right)^2 & \cdot f^{\mathrm{tr}} \quad \text{wenn} \quad \left|E^{\mathrm{P}}\right| - \left|S^{\mathrm{P}}\right| > l^{\mathrm{P}}/2 \\ \left(\frac{e}{nv}\right)^2 & \cdot f^{\mathrm{tr}} \quad \text{wenn} \quad \left|S^{\mathrm{P}}\right| - \left|E^{\mathrm{P}}\right| > l^{\mathrm{P}}/2 \\ \left(\frac{\min(s,e)}{nv}\right)^2 & \cdot f^{\mathrm{tr}} \quad \text{sonst} \end{cases} \tag{5.14}$$

$$c_{\mathrm{II}}^{\mathrm{tr}}\left(Prim^{\mathrm{P}}, Prim^{\mathrm{C}}\right) = \left(\frac{\left|\overline{M^{\mathrm{P}}M^{\mathrm{C}}}\right|}{nv}\right)^2 \cdot f^{\mathrm{tr}} \tag{5.15}$$

Die Kosten des Größenunterschieds erfassen die Abweichung der Primitivgröße. Für beide Berechnungswege I und II wird die size ()-Funktion genutzt, um die Größe der beiden Primitive zu vergleichen.

[3] Der Translationspunkt bezeichnet hier den Punkt auf dem Objekt, der zur Bestimmung der Translation betrachtet wird.

$$c_{\mathrm{I}}^{\mathrm{sc}}\left(Prim^{\mathrm{P}}, Prim^{\mathrm{C}}\right) = \left(\frac{\mathrm{size}\left(Prim^{\mathrm{P}}\right) - \mathrm{size}\left(Prim^{\mathrm{C}}\right)}{nv}\right)^2 \cdot f^{\mathrm{sc}} \tag{5.16}$$

Die Kosten des Richtungsunterschieds werden durch die Winkel-Differenz der Strecken zwischen Start- und Endpunkt bestimmt. Diese Teilkosten fließen in den Berechnungsweg I ein.

$$c_{\mathrm{I}}^{\mathrm{ro}}\left(Prim^{\mathrm{P}}, Prim^{\mathrm{C}}\right) = \sphericalangle\left(\overline{S^{\mathrm{P}}E^{\mathrm{P}}}, \overline{S^{\mathrm{C}}E^{\mathrm{C}}}\right)^2 \cdot f^{\mathrm{ro}} \tag{5.17}$$

Die Kosten für abweichende Form betreffen ebenfalls nur den Berechnungsweg I für Primitive, die mehr einer geraden Strecke ähneln als einem Kreis. Die Teilkosten betreffen die unterschiedliche Krümmung der beiden Primitive. Sie werden über den Abstand der beiden Halbpunkte H^{P} und H^{C} bestimmt. Die Punkte des Kandidaten-Primitivs $Prim^{\mathrm{C}}$ werden zuvor entsprechend der ermittelten Werte verschoben, rotiert und skaliert, sodass die Punkte S^{C} und E^{C} nun mit den Punkten S^{P} und E^{P} übereinstimmen (siehe Abbildung 5.8(a)). Dadurch unterscheiden sich die beiden Primitive nur noch durch eine unterschiedliche Krümmung voneinander.

$$c_{\mathrm{I}}^{\mathrm{sh}}\left(Prim^{\mathrm{P}}, Prim^{\mathrm{C}}\right) = \left(\frac{\left|\overline{H^{\mathrm{P}}H^{\mathrm{C}}}\right|}{\left|\overline{SE}\right|}\right)^2 \cdot f^{\mathrm{sh}} \tag{5.18}$$

Es hat sich gezeigt, dass es erforderlich ist, eine unterschiedliche Krümmungsrichtung durch zusätzliche Kosten zu bewerten. Liegt ein solcher Fall vor, werden die Kosten für abweichende Form verdoppelt.

Die Kosten für falsche Kreisteile betreffen den Berechnungsweg II. Die abweichenden Teile des Kreises – ob nun zu viel oder zu wenig – führen zu entsprechenden Kosten (siehe Abbildung 5.8(b)). Diese Abweichung ersetzt sowohl die Abweichung der Form als auch die Abweichung der Rotation des ersten Berechnungsweges.

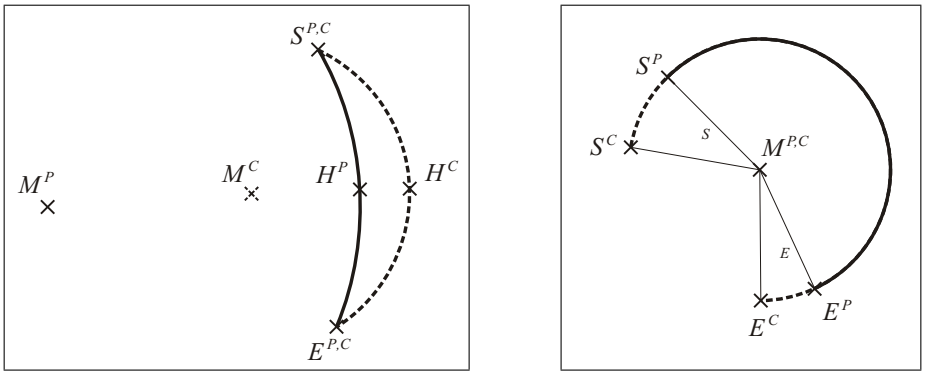
$$\delta^{\mathrm{S}} = \sphericalangle\left(\overline{M^{\mathrm{P}}S^{\mathrm{P}}}, \overline{M^{\mathrm{C}}S^{\mathrm{C}}}\right) \tag{5.19}$$

$$\delta^{\mathrm{E}} = \sphericalangle\left(\overline{M^{\mathrm{P}}E^{\mathrm{P}}}, \overline{M^{\mathrm{C}}E^{\mathrm{C}}}\right) \tag{5.20}$$

$$c_{\mathrm{II}}^{\mathrm{wp}}\left(Prim^{\mathrm{P}}, Prim^{\mathrm{C}}\right) = \left[\mathrm{abs}\left(\delta^{\mathrm{S}}\right) + \mathrm{abs}\left(\delta^{\mathrm{E}}\right)\right]^2 \cdot f^{\mathrm{wp}} \tag{5.21}$$

5.5.3 Kosten für nicht-zugewiesene Text Segmente

Die Primitive eines Prototypen repräsentieren die Striche des zu erkennenden Objekts. Während der Erkennung wird innerhalb des Kandidaten nach diesen Strichen gesucht. D. h. im Prototypen



(a) (b)

Abbildung 5.8: Kostenbestimmung (a) für abweichende Form c^{sh} zwischen zwei Primitiven mit geringer Krümmung und (b) für falsche Teile c^{wp} zwischen zwei Primitiven mit starker Krümmung.

sind Informationen über Striche gespeichert, die an einer bestimmten Stelle erwartet werden. Es gibt jedoch keine Informationen darüber, wo sich keine Striche befinden sollen. Es fehlt die Information über „verbotene Striche“. Die Teilkosten c^{us} steigen mit der Anzahl nicht genutzter Striche des Kandidaten. Es wird das Verhältnis gebildet zwischen der Länge aller Striche des Kandidaten und der Länge der Striche, die mit dem Prototypen korrespondieren:

$$c^{\mathrm{us}} = \left(1 - \frac{\textit{Länge der genutzten Striche}}{\textit{Länge aller Striche}}\right) \tag{5.22}$$

Existiert eine differenzierte Bewertung über die Zugehörigkeit der Textobjekte, wie sie in Abschnitt 3.2.3 beschrieben wurde, so kann sie an dieser Stelle genutzt werden, um unsichere Teile schwächer zu bewerten als sichere. D. h. der Längenwert wird mit einem Faktor kleiner 1 multipliziert. In diesem Fall geschah dies mit einem Wert von 0,5 für unsichere Textobjekte. Neben der zu erkennenden Schrift hängt dieser Wert vor allem von dem verwendeten Klassifizierungsalgorithmus ab.

5.5.4 Bewertung von Position und Größe der Ziffer

Bei schwierigen Schriftsituationen, wie sie des öfteren in eng geschriebenen Kirchenbüchern auftreten, können Ober- oder Unterlängen benachbarter Zeilen zu Strichkonstellationen führen, die fälschlicherweise als Ziffer erkannt werden. Um das zu verhindern, können bestimmte Annahmen über die zu erwartende Position und Größe einer Ziffer als Teil einer Textzeile getroffen werden. Für die Berechnungen werden zunächst der Abstand der Mittellinie von der Basislinie $\mathrm{dist}\,(ML, BL)$ sowie das umschreibende Rechteck mit $x_{\min}$, $x_{\max}$, $y_{\min}$, $y_{\max}$ der korrespondierenden Kandidatenprimitive bestimmt. Die Basislinie markiert hierbei die Höhe $y_{\mathrm{BL}} = 0$.

$\mathrm{pos}\,()$ ist das Produkt der einzelnen Teil-Funktionen, deren Wert größer eins wird, wenn die Ziffer zu klein ($\mathrm{pos}_1\,()$), zu hoch ($\mathrm{pos}_2\,()$) oder zu breit ist ($\mathrm{pos}_3\,()$). Solange eine Bedingung eingehalten wird, ist der Wert der jeweiligen Teilfunktion gleich eins. Ansonsten steigt der Wert linear zur Abweichung.

$$\mathrm{pos}\,() = \mathrm{pos}_1\,() \cdot \mathrm{pos}_2\,() \cdot \mathrm{pos}_3\,() \cdot \mathrm{pos}_4\,() \tag{5.23}$$

Ziffer ist zu klein. Ist die Höhe der Ziffer ($y_{max} - y_{min}$) kleiner als die Höhe der Mittellinie y_{ML}, dann steigt $\mathrm{pos}_1\,()$ linear an.

$$\mathrm{pos}_1\,(\mathrm{m}, BL, ML) = \begin{cases} 1 & \text{, wenn } y_{max} - y_{min} < y_{ML} \\ 3 - \frac{2(y_{max} - y_{min})}{y_{ML}} & \text{, sonst} \end{cases} \tag{5.24}$$

Obere Grenze der Ziffer zu hoch. Ist y_{max} größer als 1,5 mal y_{ML}, dann steigt $\mathrm{pos}_2\,()$ linear an.

$$\mathrm{pos}_2\,(\mathrm{m}, BL, ML) = \begin{cases} 1 & \text{, wenn } y_{max} < \frac{3}{2} y_{ML} \\ y_{max}/y_{ML} - \frac{1}{2} & \text{, sonst} \end{cases} \tag{5.25}$$

Obere Grenze der Ziffer zu tief. Liegt y_{max} unterhalb der Basislinie, dann steigt $\mathrm{pos}_3\,()$ linear an.

$$\mathrm{pos}_3\,(\mathrm{m}, BL, ML) = \begin{cases} 1 & \text{, wenn } y_{max} > 0 \\ 1 - y_{max}/y_{ML} & \text{, sonst} \end{cases} \tag{5.26}$$

Ziffer zu breit. Ist die Breite $x_{max} - x_{min}$ größer als 1,5 mal y_{ML}, dann steigt $\mathrm{pos}_4\,()$ linear an.

$$\mathrm{pos}_4\,(\mathrm{m}, BL, ML) = \begin{cases} 1 & \text{, wenn } x_{max} - x_{min} < \frac{3}{2} y_{ML} \\ \frac{(x_{max} - x_{min})}{y_{ML}} - \frac{1}{2} & \text{, sonst} \end{cases} \tag{5.27}$$

Die untere Grenze y_{min} der gefundenen Primitive kann nicht bewertet werden, da beispielsweise die Ziffer 9 weit unter die Basislinie reichen kann. Hier sind weitere Regelungen für spezielle Schriftstile denkbar.

Kapitel 6

Erkennung von Wörtern

Historiker sind Menschen,
die sich für die Zukunft erst interessieren,
wenn sie Vergangenheit geworden ist.
Graham Greene (1904-1991)

Das hier vorgestellte Verfahren zur Worterkennung basiert auf der im Kapitel 5 beschriebenen Ziffernerkennung. D. h. es wird ebenso eine approximative Repräsentation erzeugt und ein Matching mit Prototypen durchgeführt. Das Verfahren wurde entsprechend der gegebenen Merkmale angepasst und erweitert, wie z. B. die Einteilung des Wortes in relevante Abschnitte. Der Erkenner wurde als holistisches Verfahren entwickelt, da dieser Ansatz im Allgemeinen mit einer höheren Robustheit verbunden ist. Dies ist möglich, da bei dem hier betrachteten Problem – der Erkennung von Monatsnamen – das Lexikon abgeschlossen ist. Holistische Verfahren zur Worterkennung sind oft dadurch gekennzeichnet, dass nicht Merkmale einzelner Zeichen betrachtet werden, sondern Merkmale, die das gesamte Wort betreffen. Neben Schleifen sind es besonders die Existenz und Position von Ober- und Unterlängen [9, 83]. Neben der manuellen Erzeugung oder Anpassung der Prototypen besteht die Möglichkeit, Prototypen automatisch zu generieren. Dabei werden diese Erkenntnisse genutzt und an entsprechenden Stellen Merkmale eines Wortes extrahiert.

6.1 Aufbau der Prototypen

Die Prototypen zur Erkennung von Wörtern bestehen ebenso wie die Prototypen der Ziffernerkennung aus einer Anordnung von Primitiven. Sie werden genutzt, um innerhalb der Striche des Kandidaten nach markanten Anordnungen zu suchen; genauer gesagt: innerhalb der Primitive der approximativen Repräsentation des Kandidaten. Während maximal vier Primitive ausreichen, um eine Ziffer zu beschreiben, ist dies für ganze Wörter zu wenig. Auch wenn es nicht erforderlich ist, dass alle Striche eines Wortes repräsentiert werden, sind es doch zumindest bestimmte Stellen eines Wortes, die zur holistischen Erkennung erfasst werden müssen. Ein Prototyp für ein Wort besteht aus

mehreren Abschnitten, wobei jeder Abschnitt ein bis vier Primitive beinhaltet und eine markante Anordnung von Strichen nachbildet. Pro Abschnitt wird ein Primitiv als Hauptprimitiv definiert. Sollte eine entsprechende Bewertung möglich sein, repräsentiert dieses Primitiv einen besonders stabilen und markanten Strich eines Abschnitts. In den meisten Fällen entsprechen die Abschnitte einzelnen Buchstaben. Dies muss aber nicht zwingend so sein. Nicht selten entstehen durch zwei benachbarte Buchstaben in einem Wort Striche, die als stabiles Merkmal angesehen werden können. Vor allem bei der automatischen Generierung von Prototypen können solche Striche erfasst werden.

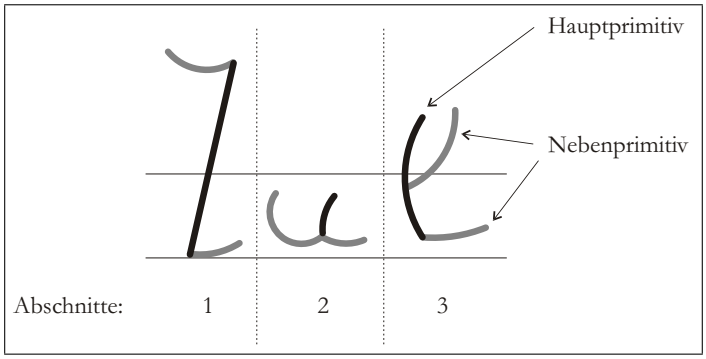


Abbildung 6.1: Beispiel eines Prototypen mit drei Abschnitten (zweiter deutscher Prototyp des Wortes Juli). Ein Abschnitt besteht aus maximal vier Primitiven. Ein Primitiv im Abschnitt ist das Hauptprimitiv, das die Position des Abschnitts bestimmt.

Für das hier vorgestellte Verfahren muss – was die Auswahl der Merkmale betrifft – zwischen der Erzeugung der Prototypen und der Erkennung unterschieden werden. Neben der manuellen Prototyp-Erzeugung gibt es die Möglichkeit, aus wenigen Beispielen eines Wortes automatisch einen Prototypen zu erstellen. Das Verfahren sucht dabei an bestimmten Stellen nach stabilen Merkmalen. Diese Stellen sind neben dem Wortanfang die Ober- und Unterlängen des Wortes. Für die Erkennung kann nicht vorausgesetzt werden, dass diese Ober- und Unterlängen auftreten. Zum Einen gibt es Handschriften, die keine ausgeprägten Ober- oder Unterlängen besitzen. Zum Anderen kann der rekonstruierte Verlauf der Textlinien ungenau sein und damit eine Ober- oder Unterlänge an einer bestimmten Position nicht gefunden werden, bzw. es werden Ober- oder Unterlängen an Stellen registriert, an denen es keine gibt. Daher wird nicht das Vorhandensein einer Ober- oder Unterlänge an einer bestimmten Stelle bewertet, sondern es wird nach Strichen gesucht, die im Normalfall eine Ober- oder Unterlänge bilden. In einem der oben beschriebenen Fälle wäre diese Ober- / Unterlänge als Merkmal nicht vorhanden, während der entsprechende Strich in etwas verkürzter Form existiert.

Die Extraktion der Merkmale konzentriert sich neben den Stellen mit Ober- und Unterlängen auf den Wortanfang in dem Bereich der ersten drei bis vier Buchstaben. Oftmals unterscheiden sich die Monatsnamen nur in diesem Bereich voneinander. Beispielsweise ist die Endung der Monatsnamen „Januar“ und „Februar“ für die letzten drei Buchstaben identisch. Desweiteren werden die Monatsnamen nicht selten abgekürzt und so können beide Varianten – abgekürzt und ausgeschrieben – von einem Modell erfasst werden. Dennoch kann sich die Stabilität der Erkennung erhöhen, wenn

auch der übrige Teil des Wortes genutzt wird – soweit dieser vorhanden ist. Denn die Unterschiede zwischen dem ausgeschriebenen Wort „Januar“ und dem Wort „Juni“ sind innerhalb der ersten drei Buchstaben marginal. Die Betrachtung der übrigen Buchstaben ist in diesem Fall sehr hilfreich.

In der aktuellen Umsetzung existieren drei Möglichkeiten, Prototypen zu erzeugen:

- das manuelle Erzeugen eines Prototypen,
- das automatische Erzeugen eines Prototypen,
- das Anpassen eines Prototypen an eine neuen Schriftstil.

6.2 Manuelle Erzeugung der Prototypen

Mittels manueller Konstruktion der Prototypen kann ein Nutzer sein Wissen über die vorliegende Schriftform einbringen, anstatt ein Training mit tausenden von Daten durchzuführen. Dies kann zwar sehr zeitaufwendig sein, aber dadurch entstehen meist Prototypen mit einem guten generischen Charakter.

Ein Prototyp wird durch das Anordnen von Primitiven gebildet, die wichtige Striche eines Wortes repräsentieren. Darin unterscheidet sich dieses Verfahren nicht von dem zuvor beschriebenen Ziffernerkenner. Ein entscheidender Unterschied ist jedoch die Gruppierung zu Abschnitten. Für die manuelle Konstruktion der Prototypen bietet es sich an, die Abschnitte so zu wählen, dass sie einzelnen Buchstaben entsprechen. Dies eröffnet die Möglichkeit, Prototypen für weitere Wörter durch die Nutzung bestehender Abschnitte zu bilden.

Mit Hilfe der so erzeugten Prototypen kann eine Erkennung von Beispielworten durchgeführt werden mit dem Zweck, die Prototypen zu überprüfen. Zum Einen werden die entstehenden Kosten für jeden Abschnitt und jedes Primitiv ausgegeben, zum Anderen kann das korrekte Matching visuell kontrolliert werden. Dadurch bekommt der Nutzer die Möglichkeit, Korrekturen durchzuführen und den Prototypen anzupassen.

6.3 Automatische Erzeugung der Prototypen

Stehen keine Experten zur Verfügung, können neue Prototypen automatisch erzeugt werden. Der Nutzer hat dabei lediglich die Aufgabe, zwei, drei oder vier typische und nach Möglichkeit vollständige Beispiele eines Wortes auszuwählen. Das System sucht dann selbstständig nach stabilen Merkmalen in allen Kandidaten.

Während des als analytisches Lernen bezeichneten Vorgangs werden aus Beobachtungen Hypothesen abgeleitet wobei vorgegebenes Wissen genutzt wird [63]. Durch die Analyse von Trainingsbeispielen

wird eine Klasse gebildet, die ähnliche Fälle in sich vereint. Erreicht wird dies durch Prototypen, die aus wenigen Beobachtungen in Form von Beispielen erzeugt werden.

Es existieren Unterschiede zwischen Instanzen unterschiedlicher Wortklassen wie auch Unterschiede der Instanzen innerhalb einer Wortklasse. Erstere Unterschiede gilt es gegenüber letzteren abzugrenzen. Für das hier dargestellte Prototyping wird von der Annahme ausgegangen, dass durch das Vergleichen von mehreren Instanzen einer Klasse stabile Merkmale gefunden werden können, die nicht nur eine bestimmte Instanz charakterisieren. Durch das Kombinieren mehrerer dieser Merkmale steigt die Wahrscheinlichkeit, dass dadurch eine Unterscheidung zwischen den Klassen möglich wird. Der resultierende Prototyp repräsentiert die Hypothese, dass mit Hilfe bestimmter Striche, die durch Primitive repräsentiert werden, das entsprechende Wort von anderen unterschieden werden kann.

Das Wissen, das mit in diesen Prozess eingebracht wird, sind Regeln über die Auswahl der in Frage kommenden Striche. In einem ersten Schritt sind dies Striche, die zu Ober- oder Unterlängen führen. Diese langen vertikalen Striche werden somit zu Kandidaten für das Hauptprimitiv eines Abschnitts, denn sie sind die stabilsten Merkmale in der Handschrift [49].

Der Algorithmus unterteilt sich in folgende Schritte:

1. Auswahl von zwei bis vier typischen Beispielen eines Wortes durch den Nutzer
2. Erzeugen der approximativen Repräsentation aller Beispiele
3. Suche nach Ober- und Unterlängen in allen Beispielen
4. Zuordnung der gefundenen Ober- und Unterlängen zwischen den Beispielen
5. Suche nach weiteren Abschnitten neben dem ersten Buchstaben
6. Suche abschnittsweise nach bester Korrespondenz von Primitiven zwischen den Beispielen

Die Art und Weise, wie nach bestimmten Abschnitten gesucht wird, hängt von der gegebenen Aufgabe ab. Für den hier betrachteten Fall der Erkennung der Monatsnamen wurde der entsprechend relevante Bereich untersucht, indem Schritt 5 durchgeführt wurde.

Die resultierenden Abschnitte werden in vier unterschiedliche Typen unterteilt:

- Oberlänge
- Unterlänge
- Kombination aus Ober- und Unterlänge
- Weder Ober- noch Unterlänge

Zunächst erfolgt somit die Suche und eindeutige Zuweisung der Abschnitte in jedem Wort. Anschließend wird in jedem Abschnitt nach einer Anordnung aus maximal drei Primitiven gesucht, die

in ähnlicher Weise in allen Beispielen vorkommt. Die Primitive des Prototypen sind das Ergebnis der Mittelung der gefundenen korrespondierenden Primitiven der Beispiele.

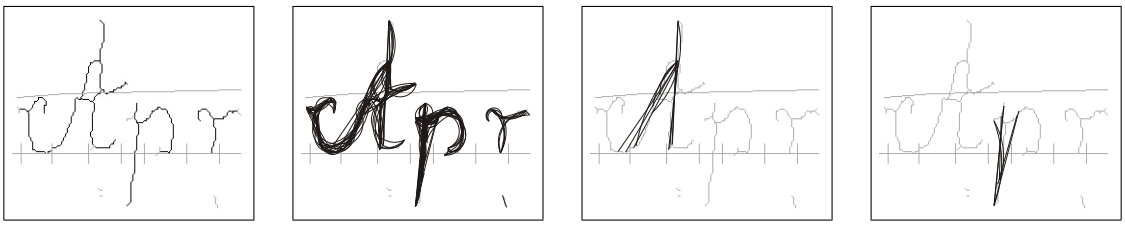
6.3.1 Suche nach Ober- und Unterlängen

Für die korrekte Funktion des Verfahrens ist eine ungefähre Angabe der Schriftgröße über die Werte der mittleren Zeichenbreite w_{Zei} und der Minuskelhöhe h_{Min} hilfreich.

Es besteht die Aufgabe, sämtliche Ober- und Unterlängen innerhalb der Beispielwörter zu finden. Zur Findung der Oberlängen werden dazu diejenigen Primitive der Approximation bestimmt, die die Mittellinie schneiden und dabei sowohl nach oben über die Mittellinie hinaus als auch nach unten in den Zeilenhauptraum eine gewisse Mindestlänge aufweisen. D. h. das vertikale Minimum des Primitivs muss dabei kleiner sein als $^2\!/_3 h_{\mathrm{Min}}$ und das vertikale Maximum größer als $^4\!/_3 h_{\mathrm{Min}}$. Diese Einschränkung stellt sicher, das nur die wahren Oberlängen erfasst werden, denn es kann durchaus vorkommen, dass auch Striche von Buchstaben ohne Oberlänge die Mittellinie schneiden. Dies liegt einerseits an der unregelmäßigen Schriftart und andererseits an einer ungenau verlaufenden rekonstruierten Mittellinie.

Die so gefundenen Primitive werden zu Oberlängen zusammengefasst (siehe Abbildung 6.2(c)). Die horizontal sortierten Schnittpunkte der Primitive mit der Mittellinie werden durchlaufen. Sobald der Abstand zum nächsten Schnittpunkt größer ist als $^3\!/_2 w_{\mathrm{Zei}}$ wird eine neue Oberlänge erzeugt.

In gleicher Weise wird beim Finden der Unterlängen verfahren (siehe Abbildung 6.2(d)). Es ist hierbei die Basislinie, die von den Primitiven geschnitten wird.



(a) (b) (c) (d)

Abbildung 6.2: Bestimmen der Primitive der Ober- und Unterlängen. (a) Skelett eines Kandidaten. (b) Approximative Repräsentation. (c) Primitive der ersten Oberlänge. (d) Primitive der ersten Unterlänge.

6.3.2 Sonderfälle der ersten Oberlänge

An dieser Stelle ergeben sich zwei kleine aber wichtige Sonderregelungen, die notwendig sind, um das Verfahren für jedes Wort funktionsfähig zu machen. Die erste Sonderregelung bezieht sich auf Worte, die mit einem großen „S“ beginnen. Die zweite Ausnahme betrifft Worte mit Anfangsbuchstaben,

deren vertikale Striche einen relativ großen horizontalen Abstand besitzen. In den hier zu Verfügung stehenden Daten betrifft das den Oktober mit seinem großen „O“.

Sonderfall „S“. Im Unterschied zu allen anderen Anfangsbuchstaben ist das große „S“ dadurch gekennzeichnet, dass es in der Regel keine einfach gebogene oder gerade Line enthält, die über die gesamte Höhe des Buchstabens verläuft. Eben diese Linien werden in dem oben beschriebenen Verfahren zum Auffinden von Oberlängen genutzt. Anstatt eines durchgehenden vertikal verlaufenden Primitivs wird daher die Kombination von zwei übereinanderliegenden Primitiven gesucht, wobei das obere eine Krümmung nach rechts und das untere eine Krümmung nach links aufweisen muss.

Sonderfall „O“. Nach dem beschriebenen Verfahren wird bei einigen Kandidaten des Monats Oktober das große „O“ als zwei Oberlängen identifiziert. Um dies zu verhindern, werden die beiden ersten Oberlängen zusammengefasst, wenn bestimmte Bedingungen eingehalten werden. Zum Einen ist dies das Vorhandensein von Primitiven, die in der ersten Oberlänge eine linksseitige und in der zweiten Oberlänge eine rechtsseitige Auswölbung besitzen, so dass sie wie eine sich öffnende und eine sich schließende Klammer ein komplettes „O“ bilden. Zum Anderen darf der Abstand zwischen diesen beiden Oberlängen-Primitiven nicht größer sein als der mittlere Abstand der potentiellen Wortgrenzen. Möglicherweise kann bei anderen, breiteren Schriften diese Sonderbetrachtung auch bei anderen Anfangsbuchstaben wie dem „M“ erforderlich sein. In den hier genutzten Daten trat dies nicht auf.

6.3.3 Zuordnung der Ober- und Unterlängen der Beispieldaten

Im Allgemeinen werden bei allen Beispieldaten die gleichen Ober- und Unterlängen gefunden. Sind unter den Trainingsdaten sowohl ausgeschriebene als auch abgekürzte Formen des Wortes vorhanden, können Ober- / Unterlängen existieren, die nicht in allen Daten vorkommen. Ein gültiger Prototyp kann nur erzeugt werden, wenn eine fehlerfreie Zuordnung der vorhandenen Merkmale stattfindet. Die Zuordnung erfolgt anhand der Positionen im Wort. Kann ein Merkmal nicht in allen Daten gefunden werden, wird es nicht weiter betrachtet.

6.3.4 Suche nach weiteren Merkmalen

Für eine vollständige Unterscheidung der Monate reichen die Merkmale der Ober- und Unterlängen nicht aus. Als Beispiel sei hier die abgekürzte Form des Januars und des Junis genannt. Weitere Merkmale sind erforderlich, auch um die Stabilität zu erhöhen. Da wichtige Merkmale innerhalb der ersten drei Buchstaben zu finden sind, wird dieser Bereich zur Merkmalsfindung untersucht.

Es kann davon ausgegangen werden, dass zuvor der erste Buchstabe als Oberlänge oder Ober- und Unterlänge erfasst wurde. Durch Nutzung der Information über die Schriftbreite w_{Zei} werden in den

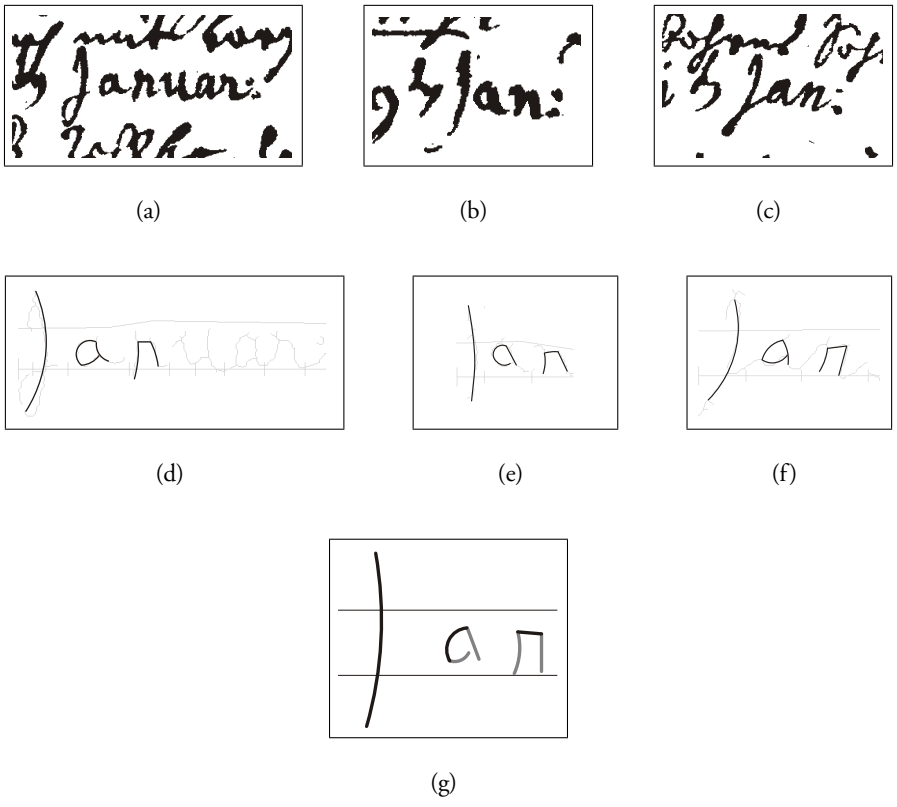
Wortbeispielen zwei Abschnitte markiert, die ungefähr dem zweiten und dritten Zeichen entsprechen. Wurde einer dieser Abschnitte als Ober- oder Unterlänge bereits erfasst, wird dieser Bereich nicht noch einmal untersucht. Dies trifft beispielsweise für das Wort „Februar“ zu, da das „b“ mit seiner Oberlänge bereits zuvor als Abschnitt untersucht wurde.

Das Ergebnis der Abschnittssuche sind Verknüpfungen von Abschnitten und Primitiven dieser Abschnitte zwischen allen Beispieldaten.

6.3.5 Finden der besten Primitivkombination

Jedes Beispiel besteht aus einer Anordnung von Strichen. Jeder Strich kann als Merkmal angesehen werden. Es werden die gemeinsamen Merkmale der Beispiele gesucht, da sie mit hoher Wahrscheinlichkeit auch Merkmale der gesamten Klasse darstellen. Dabei wird Wissen über markante Positionen angewandt, um die Suche entsprechend einzuschränken.

Die Suche erfolgt abschnittsweise. Zur Erzeugung des Prototypen wird für jeden Abschnitt die Primitivkombination ermittelt, die in allen Beispielen in möglichst ähnlicher Form zu finden ist. Pro



(a) (b) (c)

(d) (e) (f)

(g)

Abbildung 6.3: Beispiel einer automatischen Prototyp-Generierung aus drei Beispielen eines Wortes. (a)–(c) Binärbilder der Worte. (d)–(f) Korrespondierende Primitive. (g) Resultierende Primitive des Prototypen.

Abschnitt wird dabei nach einer Kombination aus maximal vier Primitiven gesucht. Existiert keine Kombination aus anfänglich vier Primitiven in allen Beispielen, so wird die Primitivzahl um eins verringert, bis eine Kombination gefunden wird, deren Kosten unterhalb von 1,0 liegen oder keine Ähnlichkeit zwischen den korrespondierenden Abschnitten festgestellt werden kann. In Abbildung 6.3 ist dies an dem Beispiel für einen Prototypen des Monats „Januar“ dargestellt. Im ersten Abschnitt konnte ein gemeinsames Primitiv gefunden werden, im zweiten und dritten Abschnitt waren es drei.

Aus den möglichen Lösungen mit ein bis vier Primitiven wird eine Lösung mit möglichst vielen Primitiven gesucht. Ein Vergleich zwischen möglichen Lösungen mit unterschiedlicher Primitiv-Anzahl erfolgt nicht. Tests, bei denen zur Entscheidung auch die Länge der gefundenen Primitive betrachtet wurde, ergaben keine Verbesserung des Verfahrens.

Die Kosten werden jeweils paarweise zwischen den Beispielen ermittelt. Um für eine Primitivkombination den größten Abstand der Beispiele im Merkmalsraum festzustellen, werden die größten Kosten zwischen den Beispielen betrachtet.

Als Beispiel seien die drei Vertreter (a), (b) und (c) des Wortes „Januar“ in Abbildung 6.3 genannt, die sowohl abgekürzt als auch ausgeschrieben auftreten. Für den ersten Abschnitt mit seinem langen Primitiv des großen „J“ werden die Kosten der Abweichung zwischen (a) und (b), zwischen (a) und (c) sowie zwischen (b) und (c) berechnet, um daraus das Maximum als Kosten für die gesamte Korrespondenz zu bestimmen.

Die Komplexität dieses Verfahrens ist sehr hoch. Angenommen, es liegen s Wortbeispiele vor. Zur Vereinfachung ist die Zahl p der Primitive in dem zu untersuchenden Abschnitt bei allen Beispielen gleich. Es soll eine Primitiv-Kombination der Größe k gefunden werden. Somit ergibt sich eine Gesamtzahl möglicher Kombinationen von $\binom{p}{k}^s$. Trotz dieser außerordentlich hohen Komplexität ist eine Nutzung dieses Verfahrens möglich, wenn die einzelnen Variablen klein gehalten werden.

Ein typisches Beispiel: Gegeben seien drei Vertreter eines Wortes. Es soll ein Abschnitt verglichen werden der jeweils zehn Primitive enthält. Die Suche einer Kombination aus vier Primitiven erzeugt somit 9 261 000 mögliche Kombinationen. Glücklicherweise ist es sehr unwahrscheinlich, dass alle Kombinationen getestet werden müssen. Sobald an einer Stelle der Berechnung die Kosten höher sind als 1,0 oder höher als das bisher ermittelte beste Ergebnisse, wird die Berechnung für diese Kombination abgebrochen. Auch ist die Zahl der Primitive, die für das Hauptprimitiv in Frage kommen, deutlich kleiner als die Gesamtzahl der Primitive eines Abschnitts. Durch solche Maßnahme lässt sich das Verfahren deutlich beschleunigen. So dauert in der Regel das Erzeugen eines Wort-Prototypen weniger als eine Minute.

6.3.6 Erzeugen der Prototyp-Primitive

Nachdem für einen Abschnitt die Suche nach der Korrespondenz erfolgreich beendet wurde, werden aus den korrespondierenden Primitiven der Beispielkandidaten die Primitive des Prototypen berechnet. Dazu werden die Parameter der Primitive einer Korrespondenz gemittelt. Somit befindet sich der Prototyp im Merkmalsraum zwischen den Beispielkandidaten.

Ein Kreisbogen kann durch unterschiedliche Merkmale beschrieben werden (siehe Kapitel 5). Genauso wie bei den beiden Wegen zur Berechnung der Kosten einer Korrespondenz zwischen zwei Primitiven in Abschnitt 5.5.2, sind auch beim Erzeugen von Prototyp-Primitiven ja nach Krümmung der Primitive die einzelnen Parameter eines Kreisbogens unterschiedlich gut geeignet. Ein Primitiv wird als kreisähnlich betrachtet, wenn der Krümmungswinkel α einen Wert von $\pi/2$ übersteigt. Ist dies bei allen Primitiven der Fall, so werden die Positionen der Kreismittelpunkte M und die Richtungen σ und ϵ der Start- und Endpunkte zum Mittelpunkt herangezogen und gemittelt. Ansonsten liegen Primitive vor, die als leicht gekrümmte Strecken betrachtet werden können. Es werden die Positionen von S, E und H zum jeweiligen Bezugspunkt bestimmt und daraus die mittleren Positionen für das neue Prototyp-Primitiv berechnet.

6.4 Anpassung der Prototypen

In Abschnitt 2.4.3 wurde erläutert, worauf die Variabilität der Schreibweise eines Wortes beruht. Während für die einzelnen *Schriftklassen* und *Schriftformen* eigene Sätze von Prototypen erzeugt werden müssen, besteht für einen neuen *Schriftstil* die Möglichkeit, einen bestehenden Satz der gleichen Schriftform anzupassen. Die Variabilität der Schreibweise eines Wortes innerhalb einer Schriftform kann in drei Kategorien unterteilt werden:

Schreibweise des Wortes. Bevor im Jahre 1902 in Deutschland die einheitliche Rechtschreibung auf Basis des 1880 erschienen Werkes von Konrad A. F. Duden beschlossen wurde, existierte keine Regelung, wie ein Wort zu schreiben war. Dadurch gibt es in Dokumenten vor dieser Zeit für ein Wort eine Mehrzahl von Schreibweisen.

Größe und Lage von Buchstaben. Zur persönlichen Ausprägung einer Handschrift gehören neben der Schriftneigung auch Eigenschaften wie die Größe der Schrift. Die Ausdehnung in horizontaler und vertikaler Richtung sowie das Verhältnis der Höhe von Minuskel und Majuskel sind Merkmale, die zur Anpassung des Verfahrens genutzt werden können.

Eigenschaften einzelner Strichklassen. In der Paläografie werden die Striche, die die Buchstaben und Wörter bilden, klassifiziert. Paläografische Aussagen über Striche einer Klasse wie beispielsweise *Schaft*, *Unterlänge* oder *Schulterstrich* für einen Schriftstil ermöglichen die Anpassung bestehender Prototypen.

Prototypen können auf drei Wegen an eine spezifische Schrift angepasst werden:

- Bearbeiten, Hinzufügen oder Entfernen von Primitiven per Hand,
- Korrektur durch Nutzung von Beispielwörtern und
- Korrektur durch Nutzung paläografischen Wissens.

6.4.1 Manuelles Bearbeiten der Prototypen

Individuelle Ausprägungen der Schrift resultieren oft in zusätzlichen Strichen oder eine eigenständige Art bestimmte Buchstaben zu schreiben. Die zusätzlichen Striche befinden sich oft am Anfang oder am Ende eines Wortes und beeinträchtigen die Erkennung dadurch, dass der Anteil an nicht zugewiesenen Strichen für alle Prototypen steigt. Andersartig geschriebene Buchstaben hingegen erfordern eine gezielte Anpassung der Prototypen.

Der strukturelle Ansatz der Worterkennung ermöglicht das Anpassen der Prototypen durch das Hinzufügen von Primitiven, die die zusätzlichen Striche repräsentieren. Desweiteren ist es möglich, Abschnitte im Prototypen entsprechend den individuellen Ausprägungen anzupassen.

Hierzu wurden Experimente durchgeführt, um die Wirkungsweise der Anpassung zu demonstrieren. Es wurden Primitive ergänzt, um die Schreibweise des Wortes „Maii“ zu korrigieren. Außerdem wurde die charakteristische Schreibung des Buchstaben „t“ im Wort „Octobre“ angepasst.

6.4.2 Automatische Anpassung der Primitiv-Parameter

Eine weitere Möglichkeit der Anpassung besteht in der Korrektur der existierenden Primitive eines Prototypen. Bei einer leichten Abweichung des Schriftstils existieren die gleichen Striche mit abweichenden Parametern wie Position oder Größe. Dies kann bei Schriften eines Schreibers über mehrere Jahre auftreten, aber auch zwischen Schriften unterschiedlicher Schreiber mit einem ähnlichen Schriftstil.

In solchen Fällen ist ein korrektes Matching zwischen Kandidat und Prototyp möglich – allerdings mit höheren Kosten, da zwar die entsprechenden Striche existieren, diese jedoch einer Abweichung unterliegen. Um diese Abweichung zu verringern und damit die Erkennung zu verbessern, werden die auftretenden Abweichungen erfasst und der Prototyp entsprechend korrigiert. D. h. es werden die Positionen der Abschnitte aktualisiert und die Parameter der Primitive dieser Abschnitte. Welche Parameter genutzt werden, hängt auch hier von deren Winkel ab (siehe dazu Abschnitt 6.3.6). Die Anpassung kann mit *einem* typischen Kandidaten erfolgen oder mit *mehreren*. Somit wird das Wissen über die stabilen Merkmale der Schriftart mit Informationen über individuelle Parameter aus den Daten kombiniert.

Es werden sämtliche Daten des Prototypen in Richtung Kandidaten verändert. In welcher Stärke dies erfolgt, wird durch einen Adaptationsfaktor $f^{\mathrm{adapt}} \in [0,1]$ bestimmt. Mit einem Wert von 0 bleibt der Prototyp unverändert, während ein Wert von 1 die maximale Anpassung an den Kandidaten bedeutet. Wie groß f^{adapt} gewählt wird, hängt von der Varianz der Schrift ab sowie von der Zahl der Kandidaten, die zur Anpassung herangezogen werden.

Sei $\overline{Merkmal}^C$ der gemittelte Wert der zur Anpassung genutzten Beispielkandidaten. Für jedes Merkmal $Merkmal^P$ eines Prototypen wird die folgende Gleichung zur Anpassung genutzt:

$$Merkmal^P_{neu} = (1 - f^{\mathrm{adapt}}) \cdot Merkmal^P_{alt} + f^{\mathrm{adapt}} \cdot \overline{Merkmal}^C \tag{6.1}$$

6.4.3 Anpassung durch Schrift-Parameter

Letztendlich besteht eine Möglichkeit, Prototypen über die Betrachtung allgemeiner Parameter eines Schriftstiles anzupassen. Die Struktur der Prototypen ermöglicht eine Klassifizierung und damit differenzierte Behandlung einzelner Primitive. Beispielsweise repräsentiert ein vertikal verlaufendes Primitiv, das über eine gewisse Mindestlänge verfügt und die Mittellinie schneidet, den *Schaft* einer Oberlänge. Ein anderes befindet sich am oberen Ende einer Oberlänge und wird dementsprechend anders klassifiziert.

Liegen entsprechende Informationen vor, so können Primitive direkt über die Art des Striches klassifiziert werden, den sie approximativ repräsentieren (*Schaft*, *Schulterstrich*, ...). Diese Informationen können für eine gezielte Anpassung der Primitive aller Prototypen eines Satzes genutzt werden. Im vorliegenden Fall wurde dieses Vorgehen anhand des Verhältnisses der Höhe der Majuskel zur Höhe der Minuskel durchgeführt.

6.5 Erzeugen der Approximation

Das Bilden der Approximation des Schriftbildes eines Monatsnamen erfolgt in der gleichen Weise wie für das Schriftbild einer Ziffer (siehe Abschnitt 5.2). In einem ersten Schritt wird aus dem Skelett eine Menge von kurzen Primitiven erzeugt, die den Verlauf der Linien nachbilden. Anschließend wird in mehreren Durchläufen versucht, aus jeweils zwei Primitiven neue, größere Primitive zu bilden – der Vorgang des sogenannten *Fusionierens*. Diese neuen Primitive werden erzeugt, wenn die Abweichung zum Verlauf der Striche einen Toleranzwert th_{P} nicht übersteigt.

Ein wesentlicher Unterschied zur Erkennung der Ziffern ist die viel höhere Zahl an Strichen und somit eine höhere Zahl an Primitiven der ersten Approximation. Für jede Approximationsstufe entspricht die Zahl der zu testenden paarweisen Kombinationen dem Quadrat der Zahl der Primitive.

Im schlechtesten Fall entstände aus jeder Kombination ein neues Primitiv und somit würde der Rechenaufwand mit jeder weiteren Approximationsstufe quadriert werden. Daher ist eine Reduzierung des Rechenaufwands unumgänglich. Erreicht wurde dies auf zwei Wegen: durch Reduzierung der Zahl der Fusionsversuche und Ausschluss bestimmter Primitive.

6.5.1 Reduzierung der Fusionsversuche

Der Toleranzwert th_{P} für ein neu zu erzeugendes Primitiv wird überschritten, wenn der räumliche Abstand zwischen den beiden Ursprungsprimitiven größer ist als $2th_{\mathrm{P}}$. Es ist nicht sinnvoll ein Primitiv, dass sich am Anfang des Wortes befindet, mit einem Primitiv am Ende des Wortes zu kombinieren.

Zunächst wird die Liste der Primitive in die Reihenfolge ihrer horizontalen Anordnung gebracht. Für ein gegebenes Primitiv $Prim_x$ werden nacheinander Kombinationen mit allen Primitiven gebildet, die rechts von $Prim_x$ liegen bis der horizontale Abstand zu groß wird. Danach gibt es keine weiteren Primitive, die mit $Prim_x$ kombiniert werden könnten. Ist der horizontale Abstand zwischen zwei Primitiven ausreichend, wird überprüft, ob der vertikale Abstand klein genug ist – ein Primitiv am Ende einer Oberlänge braucht nicht mit einem Primitiv am Ende einer Unterlänge kombiniert zu werden. Nur wenn zwei Primitive nahe genug beieinander liegen, wird eine Fusion versucht.

(a) (b)

Abbildung 6.4: (a) Beispiel einer Stelle mit vielen Kreuzungen (Wort „Dezember“ von 1813). (b) Während der Approximation werden an dieser Stelle übermäßig viele Primitive gebildet. Dies führt zu einer Erhöhung des Rechenaufwands.

Eine weitere Maßnahme zur Reduzierung des Rechenaufwands bezieht sich auf Stellen eines Wortobjektes, an denen verhältnismäßig viele Striche aufeinander treffen und viele Kreuzungen entstehen (siehe Abbildung 6.4(a)). An solchen Stellen entstehen sehr viele Primitive mit einem geringen Abstand zueinander. Beobachtungen haben gezeigt, dass während der Erzeugung der Approximation in der Regel die ersten 40 bis 50 Primitive an einer Stelle relevante Informationen zum Strichverlauf

beinhalten. Werden durch die lokale Gegebenheit weitere Primitive gebildet, erhöhen diese lediglich den Rechenaufwand. An solchen Stellen entstehen so bis zu 400 Primitive.

Daher wird für jedes Primitiv der ersten Approximation erfasst, wie oft es bereits Teil eines größeren Primitivs ist. Wird diese Zahl zu hoch, wird an dieser Stelle keine weitere Fusion durchgeführt. Die Grenze hierfür ist bei 80 festgelegt und stellt sicher, dass kein relevantes Primitiv verhindert wird.

6.5.2 Ausschluss bestimmter Primitive

Während der Approximation entstehen Primitive, die für eine weitere Fusion nicht betrachtet werden. Neben Kreisen sind dies Primitive mit einer bestimmten Position und Größe. Ein Primitiv, dass sich im Zeilenhauptraum befindet und sich horizontal über mehrere Zeichen erstreckt, ist für die Erkennung irrelevant, da es in jedem zusammenhängend geschriebenen Wort auftreten kann und keinerlei Unterscheidungsmerkmal darstellt. Gleichzeitig führt die Länge dazu, dass sich eine große Zahl von Primitiven in der Nähe befindet. Dies erhöht unnötigerweise den Rechenaufwand. Daher werden Primitive dieser Klasse nicht erzeugt.

Möglicherweise lassen sich an dieser Stelle weitere von der Schriftart abhängige Regeln aufstellen, sodass nur Primitive gebildet werden, die für die Erkennung relevante Striche repräsentieren.

Das Ergebnis einer Approximation ist in Abbildung 6.5 zu sehen. Es werden mehrere mögliche Interpretationen über den Strichverlauf durch Primitive repräsentiert. Während der Erkennung gilt es, daraus eine Untermenge von Primitiven zu finden, die möglichst gut zu den Primitiven eines Prototypen passen.

(a) Skelett des Wortes. (b) Menge der erzeugten Primitive.

Abbildung 6.5: Approximation eines Monatsnamen.

6.6 Erkennung eines Kandidaten

Auch bei der Erkennung eines Wort-Kandidaten stimmt der Ablauf mit dem der Ziffernerkenung grundsätzlich überein. Nachdem Erzeugen der Approximation des Kandidaten wird mit jedem Prototypen ein Matching durchgeführt, um zu ermitteln, für welches Wort die größten Ähnlichkeiten bestehen. Für jedes Primitiv des Prototypen wird ein korrespondierendes Primitiv des Kandidaten gesucht, dessen Kosten für die notwendigen Transformationen am geringsten sind – d. h. dessen Ähnlichkeit am größten ist.

Aufgrund der erweiterten Struktur eines Wort-Prototypen gibt es Unterschiede im Ablauf des Matchings von dem der Ziffernerkennung. Da der Prototyp aus Abschnitten besteht, wird auch bei der Erkennung abschnittsweise verfahren. Für jeden Abschnitt des Prototypen (Oberlängen, Unterlängen, Abschnitte dazwischen) wird ein entsprechendes Pendant im Kandidaten gesucht, wobei die Reihenfolge und ungefähren Abstände eingehalten werden müssen. Das Matching eines jeden Abschnittes erzeugt Kosten, die anschließend gemittelt werden. Damit ist auch die Vergleichbarkeit zwischen Prototypen mit unterschiedlicher Abschnittsanzahl gewährleistet. Eine Übersicht der auftretenden Teilkosten ist in Abbildung 6.6 dargestellt.

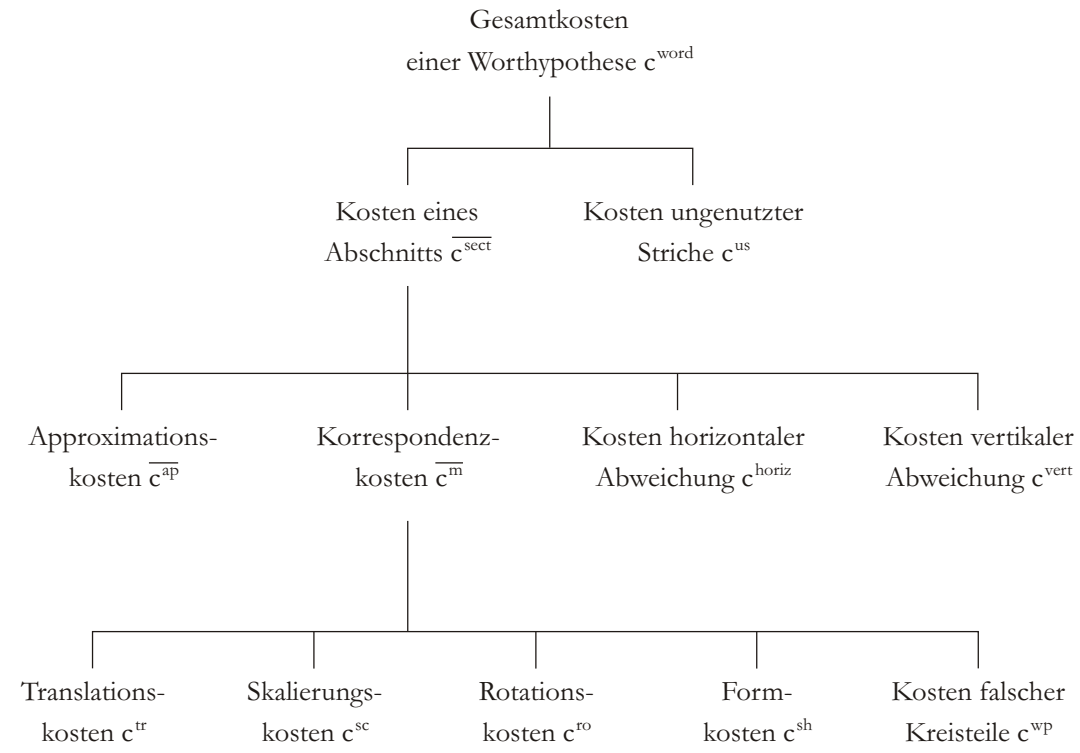


Abbildung 6.6: Übersicht über die Zusammensetzung der Kosten einer Korrespondenz zwischen Primitiven des Prototypen und Primitiven des Kandidaten eines Wortes.

6.6.1 Vergleich zwischen Prototypen

Für jeden Prototypen wird ein Matching mit den Primitiven des Kandidaten durchgeführt. Dabei entstehen Kosten – je nachdem, ob bestimmte Merkmale gut, schlecht oder gar nicht gefunden werden. Für den Vergleich zwischen diesen Korrespondenzen ist es erforderlich, den Anteil der Striche des Kandidaten, die nicht durch die Korrespondenz abgedeckt werden, als zusätzlichen Kostenfaktor zu betrachtet. Dies entspricht den Teilkosten c^{us} der Ziffernerkennung, lediglich der Faktor ist ein anderer.

Die resultierenden Gesamtkosten c^{word} für ein Matching eines Wortes setzen sich aus den Kosten c_k^{sect} mit $k = 1 \ldots T$ zusammen, die durch die Korrespondenzen der T Abschnitte entstehen (siehe nächster Abschnitt) und den Kosten der dabei ungenutzten Teile des Kandidaten:

$$c^{\mathrm{word}} = (1 - f_{\mathrm{w}}^{\mathrm{us}}) \frac{\sum_{k=1}^{T} c_k^{\mathrm{sect}}}{T} + f_{\mathrm{w}}^{\mathrm{us}} \cdot c^{\mathrm{us}} \tag{6.2}$$

Die in den Abschnitten entstehenden Kosten c^{sect} sind ebenso auf das Intervall $[0{,}1]$ normiert wie die Kosten ungenutzter Teile des Kandidaten c^{us}. Die resultierenden Kostenwerte c^{word} sind zwischen den Prototypen vergleichbar. Prototypen mit einer höheren Anzahl an Merkmalen sollen gegenüber Prototypen mit einer geringeren Merkmalszahl besser bewertet werden, wenn ein ähnlich gutes Matching vorliegt. Dies wird durch die unterschiedlichen Kostenwerte der ungenutzten Teile c^{us} erreicht. Diese Teilkosten sind bei kürzeren Prototypen entsprechend höher.

Wie in den Diagrammen in Anhang A.2 zu erkennen ist, liefert das Verfahren mit einem Faktor $f_{\mathrm{w}}^{\mathrm{us}} = 0{,}6$ gute Resultate mit allen drei Schriftstilgruppen. Die beste Leistung lässt sich jedoch erreichen, wenn der Faktor der gegebenen Schrift angepasst wird. Doch was sind die Kriterien für diese Anpassung? Die Testergebnisse lassen vermuten, dass die Güte der Schrift sowie die Existenz fremder Striche den optimalen Faktorwert beeinflussen. Unsauber geschriebene Schrift mit hohen Varianzen erfordert somit einen geringen Faktorwert ebenso wie Schriften, die durch verhältnismäßig viele fremde Striche gestört werden. Wurde der Text sauber geschrieben, so kann ein etwas höherer Wert für $f_{\mathrm{w}}^{\mathrm{us}}$ gewählt werden. Ein sinnvoller Bereich stellt hier das Intervall von 0,5 für schlechte Schriften bis 0,7 für sauber geschriebene Schriften dar.

6.6.2 Matching eines Wortes

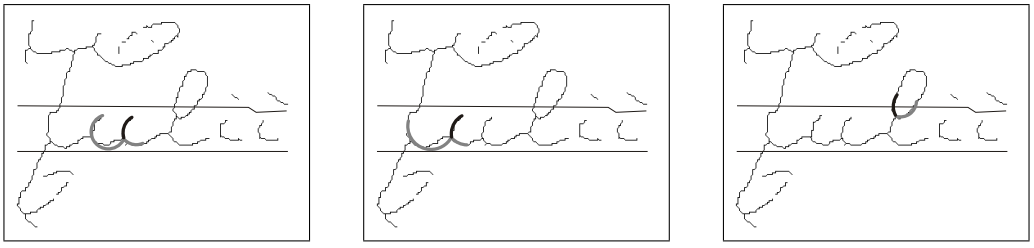
Das Matching eines Wortes erfolgt durch die Kombination von gefundenen Korrespondenzen der einzelnen Abschnitte. Für einen Abschnitt des Prototypen können mehrere potentielle Abschnitte im Kandidaten existieren.

Die Größe der Schrift ist eine der vielfältigen Variationen, die zwischen unterschiedlichen Vertretern eines Wortes auftreten können. Daher darf der horizontale Bereich, in dem ein Abschnitt gesucht wird, nicht zu stark eingeschränkt werden. So kann es vorkommen, dass in diesem Bereich mehrere potentielle Korrespondenzen für einen Abschnitt des Prototypen gefunden werden. Um das Verfahren für diese Art der Variation flexibel zu halten, werden die besten vier korrespondierenden potentiellen Abschnitte pro Prototyp-Abschnitt betrachtet. Nach den Prinzipien der dynamischen Programmierung wird aus den sich daraus ergebenen Kombinationen die beste herausgefunden. Dies erfolgt durch einen Vergleich der Kosten eines jeden potentiellen Abschnitts. Darin sind auch Kosten über die Abweichung der Position relativ zum vorhergehenden Abschnitt enthalten. Kann für einen Abschnitt des Prototypen keine Korrespondenz gefunden werden, wird ein leerer Abschnitt mit den Maximalkosten von 1,0 eingefügt.

Im Vergleich zu anderen statistisch arbeitenden Verfahren sind die hier verwendeten Strichkombinationen komplexe Merkmale. Daher ist die Zahl der in Frage kommenden potentiellen Korrespondenzen im Kandidaten gering. Die Verwendung eines Verfahrens wie des Hidden-Markow-Modells ist an dieser Stelle nicht notwendig.

Anhand eines Beispiels wird in Abbildung 6.7 dargestellt, wie mehrere potentielle korrespondierende Abschnitte im Kandidaten für einen Abschnitt des Prototypen gefunden werden.

(a) Skelett mit Textlinien. (b) Resultat der Erkennung. (c) Genutzter Prototyp.



(d) Erste Korrespondenz. (e) Zweite Korrespondenz. (f) Dritte Korrespondenz.

Abbildung 6.7: Mehrere potentielle Abschnitte werden im Kandidaten gesucht. (d)–(f) zeigen drei potentielle Korrespondenzen für den zweiten Abschnitt des Prototypen.

6.6.3 Matching eines Abschnitts

Der Abschnitt eines Prototypen besteht aus einem bis vier Primitiv(en). Im Kandidaten muss die am besten passende Kombination aus Primitiven gefunden werden. Dazu wird in einem ersten Schritt der horizontale Bereich für die Korrespondenzsuche festgelegt. Anschließend wird innerhalb der sortigen Primitive nach der besten Korrespondenz gesucht – ähnlich dem Verfahren der Zifferner-kennung.

Auch wenn es – wie oben beschrieben – eine gewisse Toleranz für die Position eines Abschnitts gibt, so kann doch der Bereich im Kandidaten eingeschränkt werden. D. h. um die Zahl der in Frage kommenden Primitive nicht unnötig zu erhöhen, wird ein horizontaler Bereich im Kandidaten für die Suche nach der besten Korrespondenz festgelegt. Dazu wird die Position des Abschnitts im Prototypen relativ zum Wortanfang bestimmt. Dabei erhöht sich die Toleranz proportional zu diesem Abstand. Für die linke und rechte Grenze x_{left} und x_{right} ergibt sich folgende Gleichung:

$$\begin{aligned} x_{left} &= 0{,}5 \cdot x_{sect} \\ x_{right} &= 1{,}5 \cdot x_{sect} \end{aligned}$$

Die hier verwendeten Faktoren haben keine sehr große Bedeutung für die Erkennung. Sie ermöglichen die Beschleunigung der Berechnung. Für die hier getesteten Schriften ist der durch diese Werte festgelegte Toleranzbereich ausreichend groß gewählt. Liegen Schriften mit geringeren Toleranzen vor, kann dieser Bereich verkleinert werden. Wird er zu klein gewählt, kann für einen Abschnitt des Prototypen der korrekte Abschnitt in einem Kandidaten nicht erfasst werden.

Die Kosten c^{sect} eines Abschnitts setzen sich aus den mittleren Matchingkosten der Primitive $\overline{c^{\mathrm{m}}}$, den mittleren Approximationskosten $\overline{c^{\mathrm{ap}}}$ sowie der Verschiebung des Abschnitts in horizontaler und vertikaler Richtung c^{horiz} und c^{vert} zusammen:

$$c^{\mathrm{sect}} = \overline{c^{\mathrm{m}}} + \overline{c^{\mathrm{ap}}} + c^{horiz} + c^{vert} \tag{6.3}$$

Die horizontale und vertikale Translation δx und δy werden auf die Größe der Schrift normiert, die durch die beiden Werte w_{Zei} und h_{Min} repräsentiert werden. Somit errechnen sich die Kosten der Translation wie folgt:

$$c^{horiz} = \left(\frac{\Delta x}{w_{\mathrm{Zei}}}\right)^2 \cdot f^{horiz} \tag{6.4}$$

$$c^{vert} = \left(\frac{\Delta y}{h_{\mathrm{Min}}}\right)^2 \cdot f^{vert} \tag{6.5}$$

Die Koeffizienten, die experimentell gefunden wurden, haben folgende Werte:

$$f^{\text{horiz}} = 0{,}2$$
$$f^{\text{vert}} = 0{,}3$$

Diese Werte können für die hier betrachteten Schriften unverändert bleiben. Das Verfahren bleibt auch bei abweichenden Einstellungen dieser Werte stabil. In Anhang A.2 sind die Zusammenhänge zwischen Parameter und Erkennungsrate für die einzelnen Schriftstilgruppen aufgeführt.

Die Positionen des Bezugspunktes im Prototypen und im Kandidaten für die Bestimmung der Translation werden durch den zuletzt gefundenen Abschnitt bestimmt. Dabei wird ein Start-/Endpunkt des Hauptprimitivs des vorherigen Abschnitts gewählt, der im Zeilenhauptraum am weitesten rechts und somit dem aktuellen Abschnitt am nächsten liegt. Wenn dieses Primitiv sowohl die Basis- als auch die Mittellinie schneidet, es sich also sowohl um eine Ober- als auch um eine Unterlänge handelt, wird der Schnittpunkt mit der Basislinie als Bezugspunkt herangezogen. Abbildung 6.8 veranschaulicht die Zusammenhänge der Bezugspunkte.

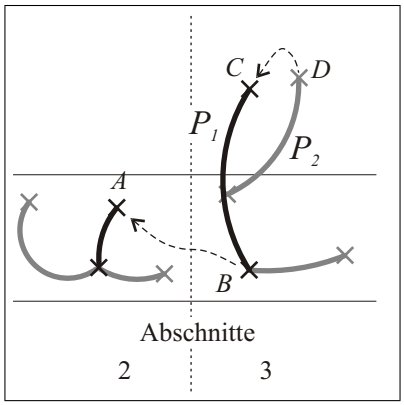


Abbildung 6.8: Detail des Prototypen „Juli“ von Abbildung 6.1 auf Seite 62. Der Bezugspunkt eines Abschnitts ist ein Punkt auf dem Hauptprimitiv des vorherigen Abschnitts. Das Hauptprimitiv P_1 des Abschnitts 3 wird durch die Position des Punktes B in Relation zu Punkt A des Hauptprimitivs des 2. Abschnitts bestimmt. Die Position des Primitivs P_2 des Abschnitts 3 wird durch die Position des Punktes D in Relation zu Punkt C bestimmt.

Kapitel 7

Ergebnisse

Verba volant, scripta manent.

Sprichwort

Dieses Kapitel befasst sich mit den Erkenntnissen und Resultaten, die aus Untersuchungen mit den zur Verfügung stehenden Daten hervorgingen. Die Ergebnisse zahlreicher Tests demonstrieren hierbei die Funktionsweise jedes einzelnen Moduls des Erkennungssystems.

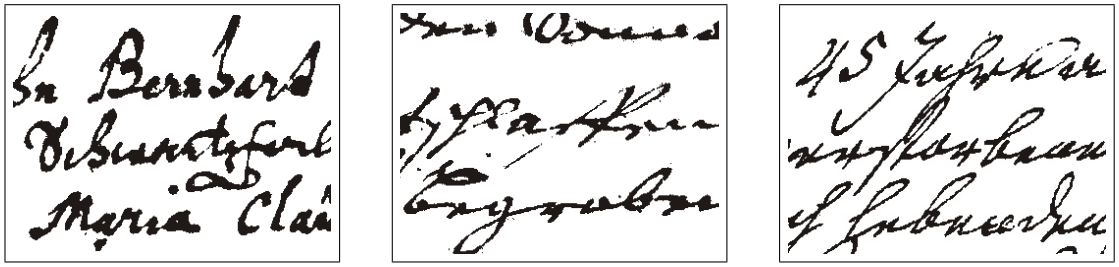
Zunächst wird noch einmal auf die Zeilensegmentierung eingegangen. In der veröffentlichten Diplomarbeit [17] wurde ein Verfahren zur Segmentierung der Zeilen in historischen Aufzeichnungen beschrieben. Es war hierbei erforderlich, eine relativ große Zahl von internen Parametern entsprechend der gegebenen Schrift einzustellen. Im Hinblick auf ein Erkennungssystem, dass sich leicht auf eine neue Schrift anpassen lässt, ist dies von Nachteil. Eine Verbesserung dieses Moduls beschränkt die Erfordernis einer Anpassung auf die Höhe und Breite der vorliegenden Schrift. Im zweiten Abschnitt werden die Tests erläutert, die zur Segmentierung des Datums durchgeführt wurden. Das Generieren der Prototypen sowie die damit erzielte Erkennungsleistung wird anschließend für die Ziffern- und die Worterkennung dargestellt. Weiterhin wird anhand von Testergebnissen gezeigt, dass mehrere Möglichkeiten der Anpassung der Prototypen eine Verbesserung der Erkennung bewirken. Schließlich wird das erfolgreiche Verarbeiten des kompletten Datums behandelt. Das Zusammenwirken der einzelnen Module sowie die Auswertung der Teilergebnisse wird beschrieben.

7.1 Segmentierung der Zeilen

Damit das Verfahren zur Segmentierung der Zeilen die bestmöglichen Resultate liefert, ist eine Reihe von Parametern auf die jeweilige Schrift einzustellen. Speziell die Suche und Bewertung potentieller Basisliniensegmente erfordert korrekte Werte.

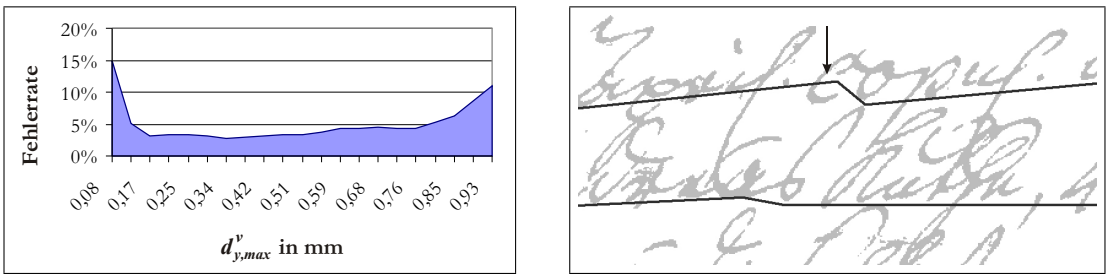
Da diese Parameter Abstandswerte in horizontaler und vertikaler Richtung repräsentieren, liegt ein direkter Zusammenhang zur Schriftgröße vor. D. h. die Parameter hängen lediglich von Merkmalen der Schrift ab, die recht einfach abgeschätzt werden können. Es sind die Höhe der Minuskel h_{Min} (entspricht dem Abstand zwischen Basis- und Mittellinie) und die mittlere Breite eines Zeichens w_{Zei} [19].

In Abbildung 7.1 sind die Werte für drei Beispielschriften aus dem 17. , 18. und 19. Jahrhundert angegeben.



$h_{\text{Min}} = 2{,}0\,mm$	$h_{\text{Min}} = 1{,}7\,mm$	$h_{\text{Min}} = 1{,}9\,mm$
$w_{\text{Zei}} = 2{,}4\,mm$	$w_{\text{Zei}} = 3{,}6\,mm$	$w_{\text{Zei}} = 3{,}0\,mm$

Abbildung 7.1: Schriftbeispiele aus den Jahren 1649, 1724 und 1812 und dessen Größenangaben.



(a) Die Abbhängigkeit der Fehlerrate von Parameter $d^{\text{v}}_{\text{y,max}}$.

(b) Beispiel eines Fehlers der rekonstruierten Basislinie.

Abbildung 7.2: Testergebnisse mit der Schrift aus dem Jahre 1812 und das Beispiel eines Fehlers.

Um den Zusammenhang zwischen Schriftgröße und korrekten Parametern festzustellen, wurden Testläufe durchgeführt, um für unterschiedliche Schriftarten die besten Einstellungen zu ermitteln. Dabei wurde ein Parameter schrittweise verändert und die resultierenden Fehler registriert. In Abbildung 7.2(a) ist dies für die maximale vertikale Abweichung $d^{\text{v}}_{\text{y,max}}$ der lokalen Minima zur Bildung der Basisliniensegmente (BLS) für Schrift aus dem Jahre 1812 dargestellt. Ein Fehler bedeutet hier eine Abweichung des Textlinienverlaufs, so dass eine falsche Unter- oder Oberlänge entsteht oder eine existierende nicht erfasst wird (siehe Abbildung 7.2(b)).

Für die Anpassung der Parameter an die jeweilige Schrift wurden die folgenden Faktoren ermittelt:

Max. horiz. Abstand der Minimumpunkte	$d^{\mathrm{v}}_{\mathrm{x,max}}$	=	$3{,}4\, w_{\mathrm{Zei}}$
Höhentoleranz der Basisliniensegmente	$d^{\mathrm{v}}_{\mathrm{y,max}}$	=	$0{,}2\, h_{\mathrm{Min}}$
Max. horiz. Abstand der BLS	$d^{\mathrm{s}}_{\mathrm{x,max}}$	=	$7{,}0\, w_{\mathrm{Zei}}$
Max. vertik. Abstand der BLS	$d^{\mathrm{s}}_{\mathrm{y,max}}$	=	$1{,}6\, h_{\mathrm{Min}}$

Die Robustheit des Algorithmus zeigt sich an der Unempfindlichkeit gegenüber Abweichungen dieser Größenangaben. In Abbildung 7.3 ist zu erkennen, dass sowohl die Schrifthöhe h_{Min} als auch die Zeichenbreite w_{Zei} relativ stark von den wahren Werten abweichen darf, ohne dass die Fehlerquote in die Höhe schnellt.

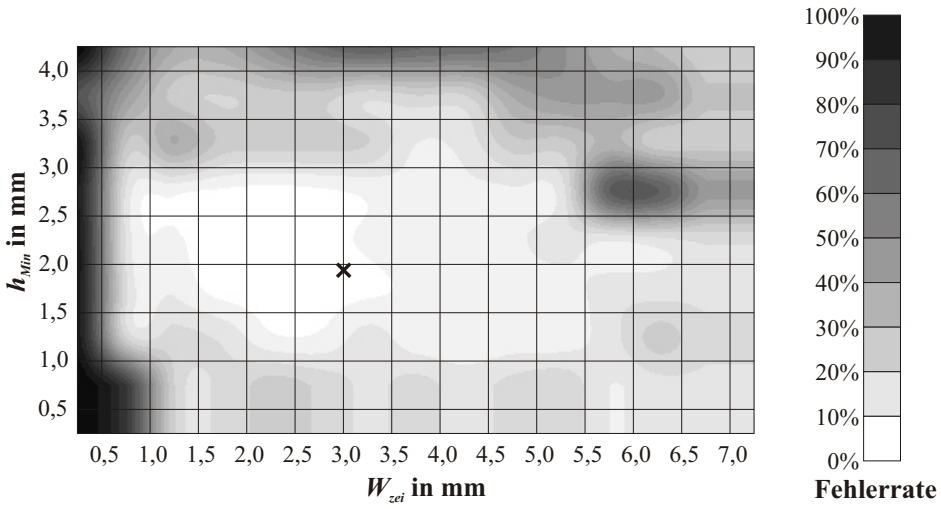


Abbildung 7.3: Abhängigkeit der Fehlerrate von den beiden anzugebenden Schriftmerkmalen h_{Min} und w_{Zei}, getestet an einem achtzeiligen Eintrag aus dem Jahre 1812. Die Position der geschätzten Werte ist durch ein Kreuz gekennzeichnet.

Wie gezeigt, basiert das Modul der Textlinienfindung und -segmentierung auf einer strukturellen Analyse und ermöglicht damit eine Anwendung ohne Training. Im Hinblick auf das Gesamtkonzept des adaptiven Erkennungssystems ist es wichtig, dass das Verfahren durch die Eingabe weniger Merkmale an einen gegebenen Text angepasst werden kann. Diese Merkmale sind Schrifthöhe und -breite und somit leicht aus der Schrift zu gewinnen.

7.2 Segmentierung des Datums

Für das Trainieren und Testen des Verfahrens zur Segmentierung des Datums wurden insgesamt 435 Einträge aus Wegenstedter Kirchenbüchern des 18. und 19. Jahrhunderts registriert. Diese enthalten insgesamt 1051 Wortgrenzen.

Auch wenn das Verfahren für alle Schriftarten und -stile unverändert angewandt wurde, werden die Ergebnisse nach den in Abschnitt 7.4 näher erläuterten Schriftstilgruppen aufgeschlüsselt. Für die Auswertung der Ergebnisse der Datumserkennung in Abschnitt 7.5 ist dies hilfreich.

Um Testläufe schnell und automatisch ablaufen lassen zu können, wurden pro Datensatz folgende Daten festgehalten:

- Angaben über Tag, Monat und Jahr
- Typ des Datums (*C-M*, *C-A-M*,...)
- Position der Wortgrenzen

Automatisch extrahiert und gespeichert wurden folgende Informationen:

- Skelett der Textobjekte, die sicher oder wahrscheinlich zum Datum gehören inklusive Strichdicke
- Verlauf der Textlinien der betreffenden Zeile sowie der darüber und darunter liegenden.

Die Zahl von 435 Einträgen ist nicht sehr groß. Für ein aussagekräftiges Ergebnis wurden daher 10 Testläufe durchgeführt, wobei durch eine zufällige Auswahl jeweils 90 % (392) als Trainings- und 10 % (43) als Testsatz genutzt wurden. Als Ergebnis lagen 92 % der korrekten Hypothesen innerhalb der besten vier Lösungen (siehe Tabelle 7.1).

Position in Ergebnisliste	Erkennungsrate in %				
	Latein 1714–1730	Latein 1768–1807	Deutsch 1807–1816	Gesamt	Akkumuliert
1	71,5	90,5	73,8	78,9	**78,9**
2	9,8	7,4	6,1	7,6	**86,4**
3	6,5	0,7	6,7	4,6	**91,0**
4	2,4	0,0	1,8	1,4	**92,4**
5	1,6	0,0	2,4	1,4	**93,8**
6	0,8	0,0	0,6	0,5	**94,3**
7	0,8	0,7	1,2	0,9	**95,2**
8	2,4	0,0	1,2	1,1	**96,3**
9	0,0	0,0	1,2	0,5	**96,8**
> 9	4,1	0,7	4,9	3,2	**100,0**
Anzahl	123	148	164	435	

Tabelle 7.1: Ergebnisse der Hypothesenbewertung der Wortgrenzen.

Der Vergleich zwischen Hypothesen unterschiedlicher Formen konnte nicht erfolgreich durchgeführt werden. Hier besteht Raum für weitere Untersuchungen. Es werden die besten Hypothesen der vier Formen für die Erkennung betrachtet.

Das Verfahren reduziert die Zahl der zu betrachtenden Hypothesen der Positionen der Wortgrenzen von mehreren tausend[1] auf weniger als zehn. Auch wenn das Potential der Analyse lokaler Merkmale noch nicht vollends ausgeschöpft scheint, so zeigen die Ergebnisse, dass dieser Ansatz der Hypothesenbildung einen Ausweg aus dem Dilemma darstellt, dass Segmentierung und Erkennung zwei sich wechselseitig bedingende Vorraussetzungen seien.

Wurden zur Bewertung der Hypothesen lediglich die lokalen Merkmale ausgewertet, so lag diese Rate der besten vier Hypothesen bei 86 %. Wurde nur die Bewertung durch die Verteilungskurven durchgeführt, waren es 67 %. Dieser Vergleich zeigt den Vorteil der Kombination der Bewertung lokaler Merkmale und a-priori-Wissen in Form von Verteilungskurven.

7.3 Erkennung der Ziffern

Auch wenn die Testdaten nicht tausende von Wortbeispielen umfassen, so besitzen diese Daten doch ein recht breites Spektrum, sodass Aussagen über die Fähigkeiten des strukturellen Erkennungsverfahrens gemacht werden können. Aus den Datumsangaben der digitalisierten Kirchenbuchseiten standen insgesamt 770 Ziffern zur Verfügung. Wie bereits in Kapitel 2 bemerkt, sind die Variationen, wie eine Ziffer geschrieben wurde, nur bedingt durch systematische paläografische Untersuchungen beschreibbar. Es überwiegen individuelle Ausprägungen.

Die zahlreichen Erscheinungsformen der Ziffern, die durch unterschiedliche Schreibweisen oder zufällige Variationen entstanden sind, werden durch mehrere Prototypen einer Ziffer abgedeckt. Die Zahl der Prototypen beträgt 3 bis 5 pro Ziffer. Die Ziffern der Wegenstedter Dokumente wurden mit durchschnittlich 80 % korrekt erkannt. Bei 89 % der Daten lag das korrekte Ergebnis innerhalb der ersten beiden Hypothesen. Siehe dazu Tabelle 7.2.

7.3.1 Parameter des Matchings zweier Primitive

Die zur Berechnung der Kosten verwendeten Faktoren dienen dazu, den Zusammenhang zwischen dem Grad der Ähnlichkeit zweier Strukturen und der Abweichung einer Eigenschaft herzustellen. Mit anderen Worten: Wie stark nimmt die Ähnlichkeit zwischen zwei Primitiven ab, wenn sie in einer Eigenschaft wie beispielsweise Ausrichtung um einen bestimmten Wert differieren?

[1] In einem Datum können bis zu 25 potentielle Grenzen vorhanden sein. Damit ergeben sich für eine Grenze 25 Hypothesen, für zwei Grenzen sind es 300 und bei drei Grenzen steigt die Zahl auf 2300 mögliche Kombinationen. Bei den vier möglichen Datumsformen ergibt das die Zahl von $25 + 300 + 300 + 2300 = 2925$.

Ziffer	Beispiele	Zahl der Prototypen	Erkennungsrate Pos. 1 in %	Erkennungsrate Pos. 1 + 2 in %
0	36	5	89	94
1	192	3	83	87
2	228	4	83	87
3	63	4	76	84
4	33	4	73	88
5	46	5	78	91
6	45	3	84	98
7	44	3	77	82
8	38	4	76	89
9	44	5	77	86

Tabelle 7.2: Testergebnis der Ziffernerkennung. Unterschiedliche Erscheinungsformen erfordern mehrere Prototypen pro Ziffer.

Mit Hilfe der Testdaten wurden diese Parameter eingestellt. Dafür wurden bei einer falschen Erkennung die entstandenen Teilkosten zwischen dem nicht korrekten aber am besten bewerteten Prototypen und dem korrekten Prototypen verglichen. Ähnlich dem Back-Propagation Verfahren wie es beispielsweise zum Trainieren eines Neuronalen Netzes genutzt wird, wurden die Faktoren derjenigen Teilkosten erhöht, die für die korrekte Lösung geringer waren als bei der besten Lösung, während die Faktoren für die höheren Teilkosten verringert wurden. Auch wenn diese Anpassung automatisiert ablaufen kann, wurde sie manuell durchgeführt, da bei Fehlleistungen abgeschätzt werden muss, inwieweit der Fehler auf die Parametrisierung oder den entsprechenden Prototypen zurückzuführen ist.

Abschließend wurden Testläufe durchgeführt, die zum Einen die optimale Einstellung verifizieren und zum Anderen die Robustheit des Verfahrens gegenüber abweichende Einstellungen der Parameter demonstrieren. In Abbildung 7.4 ist der Zusammenhang zwischen Erkennungsrate und dem Wert des Faktors $f_{\mathrm{z}}^{\mathrm{us}}$ dargestellt. Dies ist der einzige Faktor, der ein deutliches Maximum aufweist. Dieser wird auch als einziger Parameter dem jeweiligen Problem (Ziffernerkennung, Worterkennung) entsprechend angepasst. Alle weiteren Faktoren wie Rotation (f^{ro}) oder Translation (f^{tr}) können über einen großen Bereich variieren, ohne dass die Erkennungsrate signifikant abfällt. Die detaillierte Darstellung der Zusammenhänge zwischen den Faktoren und der Erkennungsrate sind in Anhang A.1 aufgeführt.

Die Faktoren wurden so skaliert, dass entweder nur Kostenwerte in einem Bereich von $[0,1]$ entstehen können oder Werte größer eins nicht weiter betrachtet werden. Sind die Kostenwerte größer, ist die Ähnlichkeit klein genug, um an dieser Stelle die entsprechende Korrelation zu verwerfen. Der

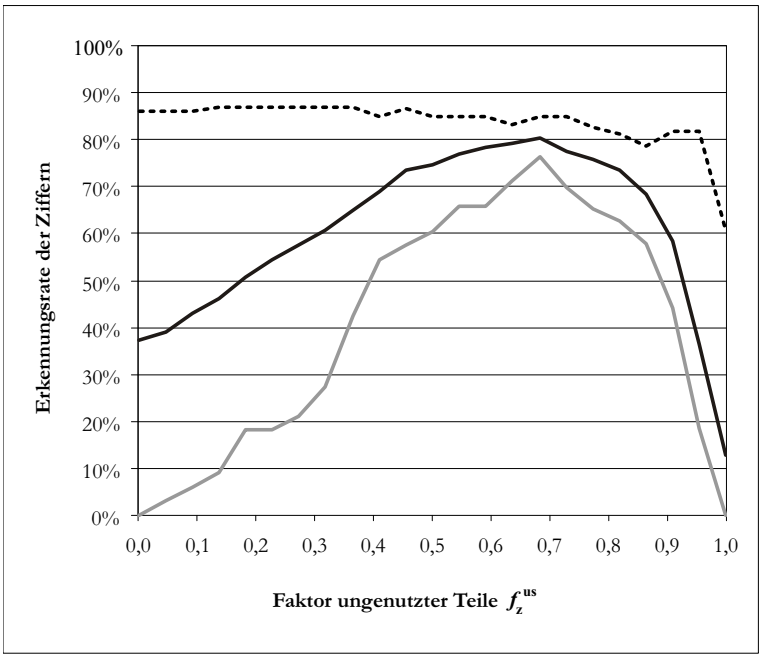


Abbildung 7.4: Verlauf der Erkennungsrate in Abhängigkeit des Kostenfaktors f_z^{us}. Die drei Kurven repräsentieren die Erkennungsrate der besten Ziffer (gestrichelt), der schlechtesten Ziffer (grau) und die mittlere Erkennungsrate (schwarz).

einzige Faktor, der sich daraus direkt ergibt, ist der Faktor der Rotationskosten f^{ro}. Die maximale Abweichung, die bei einer Rotation auftreten kann, ist $\pi/2$. In Verbindung mit Gleichung 5.17 auf Seite 58 ergibt sich somit ein Faktor von 0,4.

Für die anderen Faktoren wurden die folgenden Werte ermittelt:

$$\begin{aligned} f^{ap} &= 25{,}0 \\ f^{tr} &= 2{,}2 \\ f^{sc} &= 0{,}9 \\ f^{sh} &= 1{,}2 \\ f^{wp} &= 0{,}12 \\ f_z^{us} &= 0{,}7 \end{aligned}$$

Ein Beispiel eines konkreten Kostenwertes veranschaulicht, welche Abweichungen der einzelnen Transformationen einander entsprechen. Der Kostenwert von 0,5 entsteht wenn zwei Primitive identisch sind, nur

- die Position sich um ungefähr 50 % der Primitivgröße unterscheidet.
- ein Größenunterschied vorliegt, bei dem das kleinere Primitive 57 % der Größe des größeren Primitivs besitzt.
- ein Richtungsunterschied zwischen Typ-I-Primitiven von $64°$ vorliegt.

- ein Krümmungsunterschied zwischen Typ-I-Primitiven besteht, sodass die Halbpunkte einen Abstand von rund 1/3 der Primitivgröße haben und eine unterschiedliche Krümmungsrichtung vorliegt.
- die abweichenden Kreisteile zweier Typ-II-Primitive zusammen rund 114° betragen.

Der gleiche Wert entsteht, wenn die Abweichung des Strichverlaufs vom Primitiv 14 % der Primitivgröße entspricht.

Für jeden der hier angeführten Parameter gibt es ein Optimum aber das erfolgreiche Arbeiten des Verfahrens setzt nicht voraus, dass diese Werte exakt getroffen werden. Die Ergebnisse dieser Untersuchungen sind die Erkenntnisse, in welchem Verhältnis die einzelnen Formen der Abweichung zueinander stehen. Abgesehen von den Kosten nicht genutzter Striche sind diese gewonnenen Erkenntnisse allgemeingültig und unabhängig von dem vorliegenden Problem – Ziffern- oder Worterkennung, deutsch oder latein. Sie werden auch für die Worterkennung genutzt, deren Ergebnisse im Abschnitt 7.4 erläutert werden.

7.3.2 Fehlerquellen und Grenzen des Verfahrens

Durch die Wahl der Striche als Merkmal bietet dieser strukturelle Ansatz eine Robustheit gegenüber Störungen, die durch falsche Striche erzeugt werden und damit in ihrer Form den Merkmalsträgern ähneln. Dennoch wirkt sich diese Eigenschaft bei den Ziffern nicht so stark aus wie erwartet. Dies liegt vor allem an der vergleichsweise geringen Zahl der Merkmale. Verbunden mit der hohen Varianz der Schrift können so bei dem Vorhandensein fremder Striche benachbarter Zeilen oder Wörter die Merkmale einer anderen Ziffer entstehen (siehe Abbildung 7.5). Da die Größe einer Ziffer stark variieren kann, kann das Größenverhältnis nicht so stark eingeschränkt werden, um diesen Fehler zu verhindern.

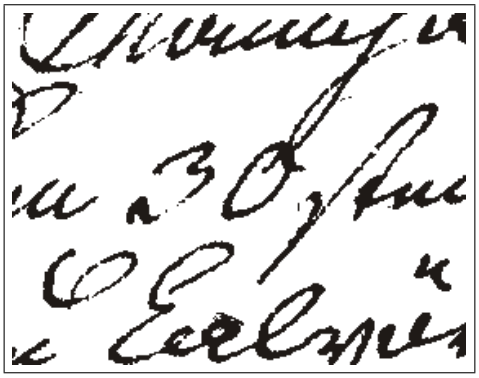


Abbildung 7.5: Durch Unterlängen darüber liegender Zeilen kann die Erkennung einer Ziffer erschwert werden. Die Folge: Eine Null kann als Acht erkannt werden. Erst die Nutzung von Regeln über gültige Datumsangaben ermöglichen die korrekte Erkennung.

Jedes Verfahren hat seine Grenzen. Und so kann auch die hier vorgestellte Ziffernerkennung falsche Ergebnisse liefern, wenn der Grad der Störung zu hoch wird. Neben den falschen Strichen gilt dies sowohl für zu viel Tinte während des Schreibvorgangs als auch für zu wenig. Im ersten Fall führt dies zu einer hohen Strichdicke oder zu Flecken, die nach der Skelettierung zu unvorhersagbaren Artefakten des Skeletts führen. Im zweiten Fall kann ein zu schwacher Strich nach der Vorverarbeitung nicht als solcher erkannt werden und fehlt als Merkmal während der Erkennung. Beide Effekte können auch auftreten, wenn einerseits Stockflecken oder andere Verunreinigungen durch die Vorverarbeitung nicht entfernt werden konnten und andererseits die Tinte über die Zeit ausgeblichen ist (siehe Abbildung 7.6).

(a) Grauwertbild. (b) Binärbild. (c) Resultat der Skelettierung.

Abbildung 7.6: Das durch Verunreinigung verdunkelte Papier führt zu Artefakten nach der Binarisierung. Dadurch entsteht ein gestörtes Skelettbild, mit der Folge, dass die Ziffer 1 nicht mehr korrekt erkannt werden konnte.

7.4 Erkennung der Monatsnamen

Im Unterschied zu Ziffern hängt die Erscheinungsform von Wörtern sehr stark von der jeweiligen Schriftform und vom Schriftstil ab. Daher ist es hilfreich, die zur Verfügung stehenden Daten hinsichtlich der Schriftformen und Schriftstile zu untersuchen und zu klassifizieren.

Von der Gemeinde Wegenstedt wurden 104 Seiten aus Kirchenbüchern des 18. und 19. Jahrhunderts digitalisiert. Daraus resultierten 435 Monatsnamen aus kompletten Datumsangaben und 16 Monatsnamen, die in einem anderen Zusammenhang in den Dokumenten auftraten. Von den insgesamt 451 wurden 271 in lateinischer und 180 in deutscher Schrift geschrieben. Zur Erweiterung dieser Datenbasis wurden 126 Monatsnamen von Kirchenbüchern der Gemeinde St. Johannes in Schönebeck / Bad Salzelmen digitalisiert. Diese stellen eine weitere lateinische Schriftstilgruppe dar. Sie wurde zur Demonstration der Anpassbarkeit bestehenden Prototypen genutzt.

Während der Sichtung der Daten wurden die Schriftproben klassifiziert. Eine Differenzierung der Daten hinsichtlich der Schriftstile fällt schwer, da sich über die einzelnen Monatsnamen hinweg kein kontinuierliches Bild abzeichnet. D. h. nicht jeder Schriftstil ist bei jedem Monatsnamen vertreten.

In einer ersten Einteilung basierend auf Zeit und Ort lassen sich vier Gruppen von Schriftstilen bilden – drei aus Wegenstedt, eine aus Schönebeck / Bad Salzelmen. Diese Einteilung ist sinnvoll, da für eine Erkennungsaufgabe eines Dokuments dessen Herkunft und Zeit herangezogen wird, um die Prototypen der in Frage kommenden Schriftstile auszuwählen. Die vier Gruppen werden im Folgenden benannt und charakterisiert, Beispiele sind in Abbildung 7.7 dargestellt.

Gruppe A, latein, Wegenstedt, 1714–1730. Diese Stilgruppe zeichnet sich durch eine allgemein große Varianz aus. Dies gilt für den Verlauf der Textlinien ebenso wie für die Ausprägung von Schleifen und Bögen eines Zeichens. Es wurde augenscheinlich ein anderes Schreibgerät genutzt als 50 Jahre später. Die Dicke der Linien schwankt und ein Zuviel an Tinte führt häufiger zu einem Zulaufen von Schleifen. Vermutlich liegen zwei unterschiedliche Schriftstile vor. Am augenfälligsten ist die Unterscheidung anhand des Vorhandenseins einer abschließenden Unterlänge. Aber auch die Schreibweise des kleinen „p“ im Wort „April“ lässt den unterschiedlichen Stil erkennen.

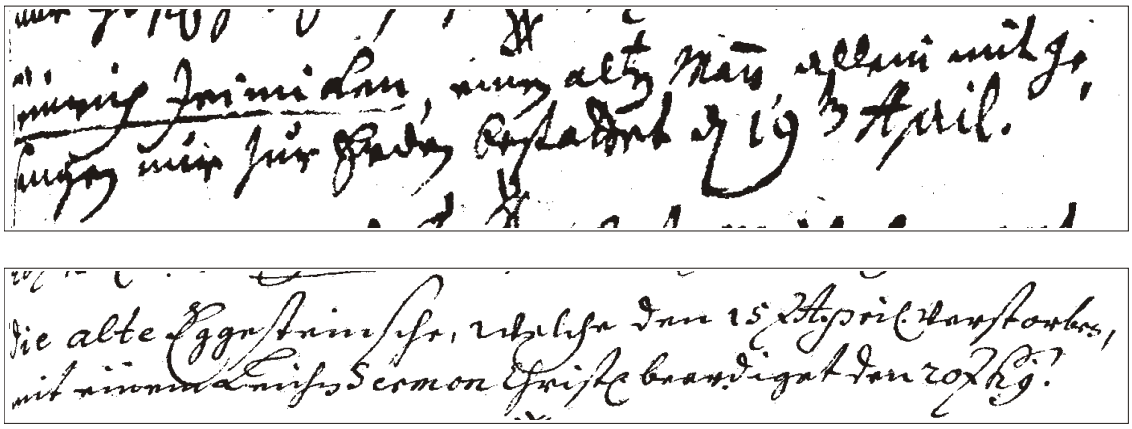
Gruppe B, latein, Wegenstedt, 1768–1807. Im Gegensatz zu Gruppe A ist diese Schrift durch Kontinuität und Sauberkeit geprägt. Hier ist nur ein Schriftstil erkennbar. Die Variationen innerhalb dieses Stils sind gering genug, dass die Erkennung mit jeweils einem Prototypen pro Wort möglich ist. Die Schrifthöhe ist geringer als bei der vorhergehenden Gruppe. Allerdings besteht hier die Besonderheit, dass die Minuskelhöhe der lateinischen Monatsnamen meist größer ist als die des ansonsten in deutsch gehaltenen Texts. Das Gleiche gilt für die Namen der beteiligten Personen, die durch Frakturschrift hervorgehoben wurden.

Gruppe C, deutsch, Wegenstedt, 1807–1816. In dieser Gruppe wurde verhältnismäßig sauber geschrieben. Allerdings führt eine geringer Zeilenabstand verbunden mit großzügigen Ober- und Unterlängen häufig zu Artefakten durch Striche benachbarter Zeilen. Die Monatsnamen wurden ausgeschrieben. Anhand der Anfangsbuchstaben einiger Monatsnamen lassen sich zwei Stile unterscheiden. So wird das große „A“ der Monate „April“ und „August“ auf zwei unterschiedliche Arten geschrieben. Das Gleiche gilt für das „S“ des „September“.

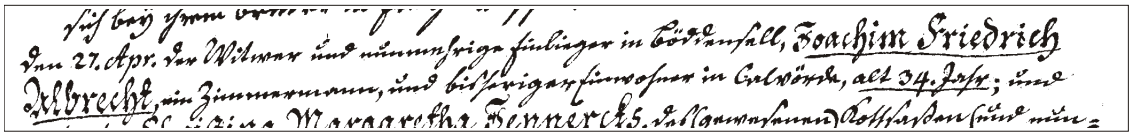
Gruppe D, latein, Schönebeck / Bad-Salzelmen, 1766–1780. Diese Schrift ähnelt dem Stil der Gruppe B. Die hier enthaltenen Schriftstile sind vor allem durch die unterschiedlichen Ausprägungen des großen Anfangsbuchstabens eines Wortes gekennzeichnet. Anders als bei Gruppe C ist die Art eines Zeichens gleich, aber zusätzliche Striche und Schleifen machen eine Anpassung erforderlich. Durch dieses Merkmal konnten drei Schriftstile erkannt werden.

Für die Erkennung der Wegenstedter Daten wurden Prototypen mit Hilfe typischer Vertreter eines Wortes automatisch generiert. Anschließend wurden manuelle Nachbesserungen an Protoypen vorgenommen, wenn erkennbar war, dass eine andere Strichkombination die Menge der Kandidaten besser repräsentierte. Die Zahl der Prototypen je Wort wurde von der Vielfalt der Schriftstile bestimmt, in denen dieses Wort auftrat. Mehrere Prototypen waren vor allem dann erforderlich, wenn

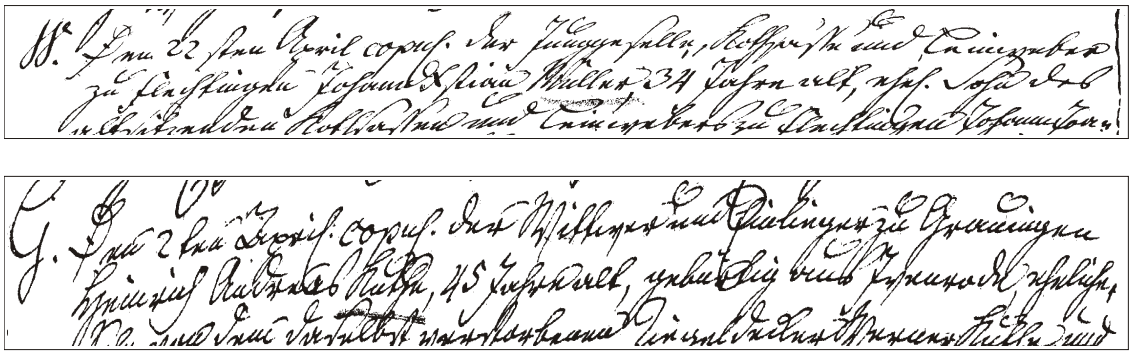
Gruppe A



Gruppe B



Gruppe C



Gruppe D

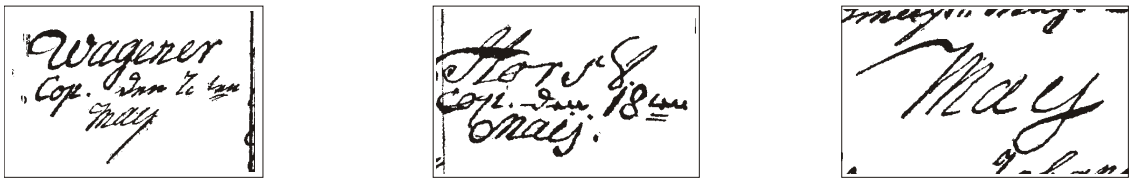


Abbildung 7.7: Übersicht über die vorliegenden Schriftstile und ihre Gruppierung.

Monat	Beispiele	Zahl der Prototypen	Erkennungsrate Pos. 1 in %	Erkennungsrate Pos. 1 + 2 in %
Januar	12	1	75	92
Februar	10	2	70	80
März	15	2	80	100
April	22	3	73	73
Mai	12	2	83	100
Juni	6	1	83	83
Juli	6	1	100	100
August	8	2	63	63
September	8	1	88	88
Oktober	5	1	100	100
November	11	2	82	95
Dezember	8	2	78	100

Tabelle 7.3: Testergebnis der Monatserkennung für Stilgruppe A (lateinisch).

Monat	Beispiele	Zahl der Prototypen	Erkennungsrate Pos. 1 in %	Erkennungsrate Pos. 1 + 2 in %
Januar	14	1	79	93
Februar	11	1	91	100
März	2	1	100	100
April	8	1	100	100
Mai	8	1	87	100
Juni	6	1	83	100
Juli	5	1	100	100
August	0	0	–	–
September	7	1	100	100
Oktober	12	1	83	92
November	61	1	93	98
Dezember	11	1	91	100

Tabelle 7.4: Testergebnis der Monatserkennung für Stilgruppe B (lateinisch).

es von mehreren Buchstaben eines Wortes mehrere Versionen gab. So waren beispielsweise für das Wort Mai in der Stilgruppe A drei Prototypen notwendig, um eine Erkannungsrate von über 80 % zu erzielen, da sowohl das große 'M' als auch das kleine 'a' in unterschiedlichen Formen auftraten – ein Nachteil des holistischen Ansatzes.

Monat	Beispiele	Zahl der Prototypen	Erkennungsrate Pos. 1 in %	Erkennungsrate Pos. 1 + 2 in %
Januar	15	2	87	100
Februar	11	2	100	100
März	20	1	80	85
April	18	2	78	100
Mai	15	2	80	80
Juni	8	2	75	88
Juli	11	2	82	91
August	9	2	89	89
September	10	2	90	100
Oktober	11	2	73	100
November	34	1	85	94
Dezember	19	1	84	90

Tabelle 7.5: Testergebnis der Monatserkennung für Stilgruppe C (deutsch).

Für die lateinische Gruppe A lag die Erkennungsrate insgesamt bei **79 %** (87 % mit Zweitplatzierten). Bei der etwas sauberer geschriebenen Schrift der Gruppe B wurde eine Rate von **93 %** erreicht (98 % mit Zweitplatzierten). Die in deutsch geschriebenen Monatsnamen der Gruppe C wurden in **84 %** (93 % mit Zweitplatzierten) korrekt erkannt. Aufgeschlüsselt für die einzelnen Monate sind die Ergebnisse in den Tabellen 7.3, 7.4 und 7.5 dargestellt.

Die erreichten Ergebnisse zeigen den Erfolg dieses Ansatzes. Die Werte liegen in dem Bereich, in dem auch die Werte vergleichbarer Verfahren zur Worterkennung liegen. Das statistische Erkennungsverfahren von Morita et al. wurde zur Erkennung von Monatsnamen auf Bank Schecks entwickelt. Nach einem Training mit 1190 Daten wurde eine Erkennungsrate von 91 % erreicht. In der von Oliveira Jr. et al. vorgestellten Arbeit [71] war das Ziel, die beste Merkmalskombination zu finden, um Monatsnamen zu erkennen. Erreicht wurde eine mittlere Erkennungsrate von 87 %. Auch im Bezug zu Verfahren, die speziell für die Verarbeitung historischer Dokumente entwickelt wurden, sind die erreichten Ergebnisse vielversprechend. Rath et al. [77] erreichte mit der Word Spotting Technik eine Genauigkeit von 72 %.

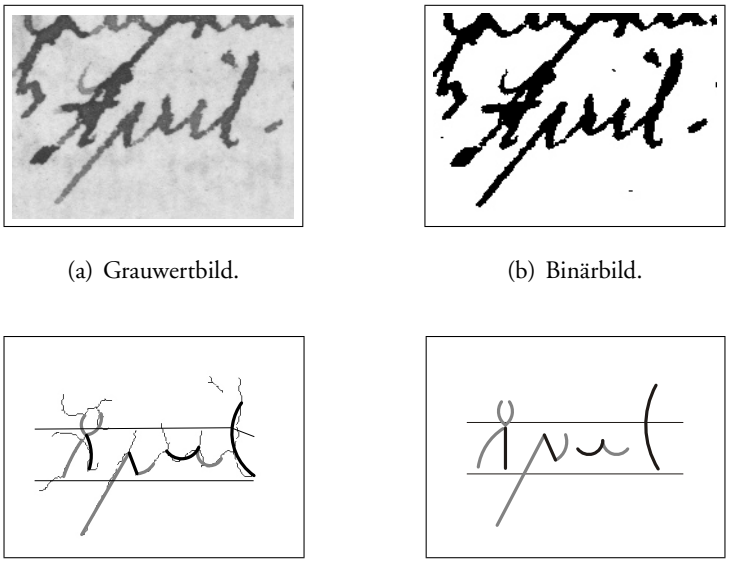
7.4.1 Robustheit

Da Robustheit ein Merkmal ist, was sich nur schwer quantitativ beschreiben lässt, werden im Folgenden Beispiele von Monatsnamen dargestellt, die trotz einer Störung durch fehlende oder fremde Striche korrekt erkannt wurden (siehe Abbildungen 7.8, 7.9, 7.10 und 7.11).

Die Störungen sind vielseitig. In den meisten Fällen resultieren Artefakte aus Unterlängen darüberliegender Zeilen. Zum Einen ragen diese in den Zeilenraum der aktuellen Zeile, zum Anderen kommt es zu Berührungen oder Überschneidungen dieser fremden Striche mit den Strichen des zu erkennenden Wortes. Im ersteren Fall werden diese Artefakte in der Regel durch die zuvor stattfindende Zeilensegmentierung erfolgreich entfernt. In den Abbildungen 7.9(b) und 7.9(c) ist zu erkennen, dass die Unterlänge des „g“ der vorherigen Zeile bis zur ersten Unterbrechung des Strichs aus dem aktuellen Schriftbild entfernt wird. Ebenso geschieht dies mit der Unterlänge am rechten Bildrand der Abbildung 7.10(b). Auch Striche, die zu einfachen Berührungen führen, werden meistens durch die Vorverarbeitung korrekt zugeordnet. Zu sehen ist dies ebenfalls in Abbildung 7.10(b), wo es zu einer Berührung zwischen dem „M“ und einer Unterlänge kommt. Sobald jedoch eine eindeutige Zuordnung eines Striches zu einer Zeile nicht mehr möglich ist, bleiben sie für den anschließenden Erkennungsprozess bestehen. Sowohl in Abbildung 7.8(b) als auch in Abbildung 7.9(b) ist dies bei dem letzten Bogen der Unterlänge der Fall.

Eine weitere Form der Störungen sind lückenhafte oder fehlende Striche. Dafür gibt es mehrere Ursachen:

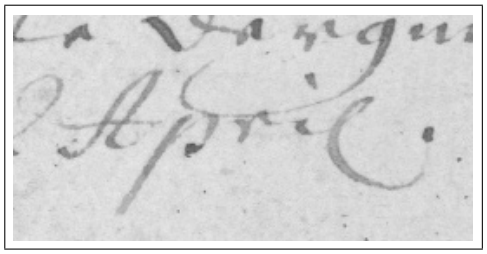
- Der Tintenfluss während des Schreibens lässt nach oder reißt ab.
- Die Tinte bleichte mit der Zeit aus.
- Das Papier ist durch Alterung oder Verschmutzung dunkel geworden.



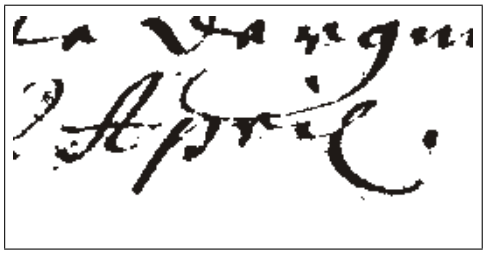
(a) Grauwertbild. (b) Binärbild.

(c) Resultat der Erkennung. (d) Genutzter Prototyp.

Abbildung 7.8: Das Wort „April“ aus einem Wegenstedter Eintrag aus dem Jahr 1721. Durch eine Überschneidung mit einer Unterlänge kommt es zu Artefakten am A.

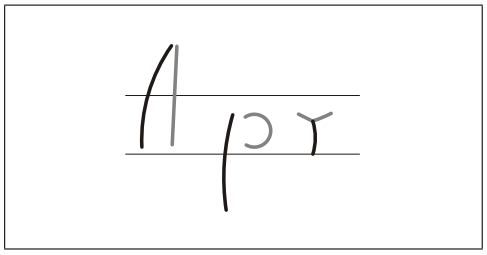


(a) Grauwertbild.



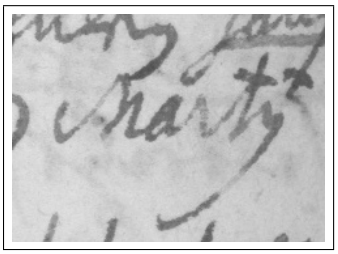
(b) Binärbild.

(c) Resultat der Erkennung.

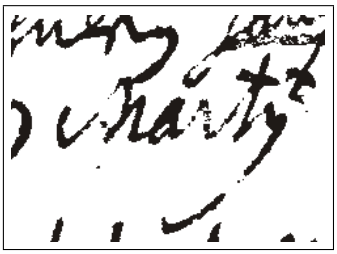


(d) Genutzter Prototyp.

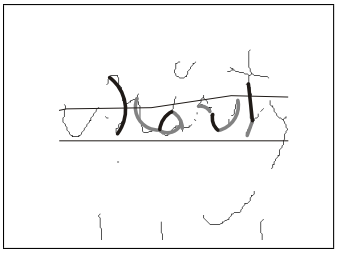
Abbildung 7.9: Wort „April“ aus einem Wegenstedter Eintrag aus dem Jahr 1723. Die Unterlänge berührt an mehreren Stellen das Wort.



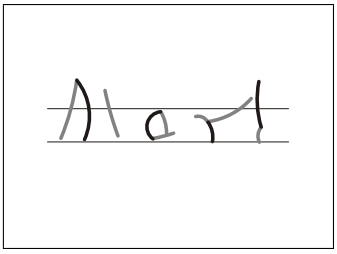
(a) Grauwertbild.



(b) Binärbild.

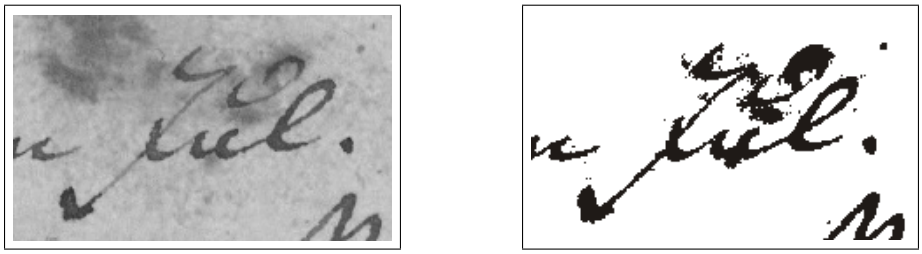


(c) Resultat der Erkennung.

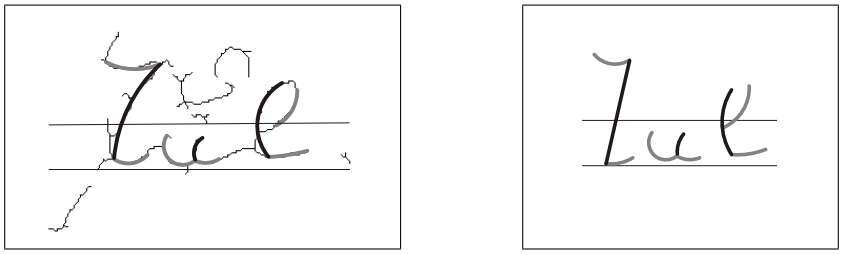


(d) Genutzter Prototyp.

Abbildung 7.10: Wort „Martij“ aus einem Wegenstedter Eintrag aus dem Jahr 1720. Das Wort konnte trotz des fehlenden Strichs im M und des lückenhaften r korrekt erkannt werden. Die Unterlängen der darüberliegenden Zeile wurden von der Zeilensegmentierung größtenteils korrekt zugeordnet.



(a) Grauwertbild. (b) Binärbild.



(c) Resultat der Erkennung. (d) Genutzter Prototyp.

Abbildung 7.11: Abgekürztes Wort „Jul“ aus einem Wegenstedter Eintrag aus dem Jahr 1807. Die durch Stockflecken verursachten Artefakte beeinflussen die Erkennung des Wortes nicht.

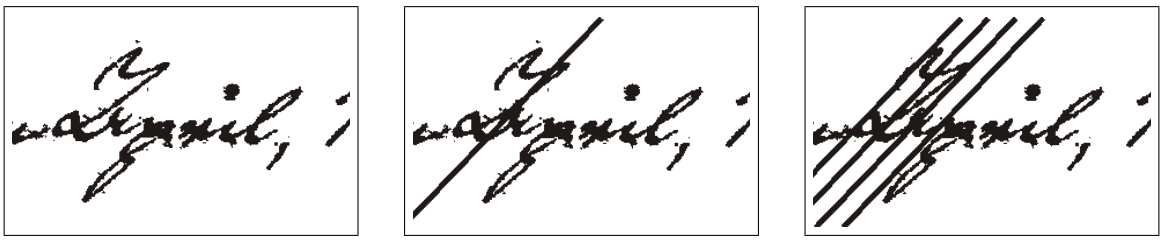
In jedem Fall ist das Ergebnis ein zu geringer Kontrast zwischen Vorder- und Hintergrund, sodass der Strich nicht erfasst werden kann. Deutlich wird dies in den Abbildungen 7.10(a) und 7.10(b). Sowohl das fehlende Strichteil beim „M“ als auch das schwach ausgeprägte „r“ erschweren die Erkennung des Wortes.

Schließlich führen Verunreinigungen wie Stockflecken zu dunklen Stellen im Papier, wodurch wiederum störende Artefakte für die Erkennung auftreten. Der Verlauf der Striche kann an diesen Stellen nicht fehlerfrei erkannt werden. Als Beispiel ist hier das abgekürzte Wort des Juli in Abbildung 7.11(a) dargestellt. Es sind deutlich die dunklen Stellen im Papier zu erkennen, sodass eine Unterscheidung zwischen Schrift und Hintergrund mittels Helligkeitswert sehr schwer fällt. Das Erkennungsverfahren kann trotz dieser Artefakte das Wort korrekt erkennen.

Die am häufigsten auftretenden Artefakte werden durch Unterlängen der darüberliegenden Zeile verursacht. Um den Einfluss dieser Artefakte zu untersuchen, wurde in mehreren Durchläufen eine Erkennung eines Monatsnamen durchgeführt, der durch das Hinzufügen von Strichen gestört wurde, die von oben nach unten durch das gesamte Wort verlaufen (siehe Abbildung 7.12). In einem ersten Testlauf wurde diese Störung an unterschiedlichen Positionen getestet. In einem weiteren Durchlauf wurde die Zahl der Unterlängen erhöht. Ausgewählt wurde ein in deutsch geschriebener Monatsname.

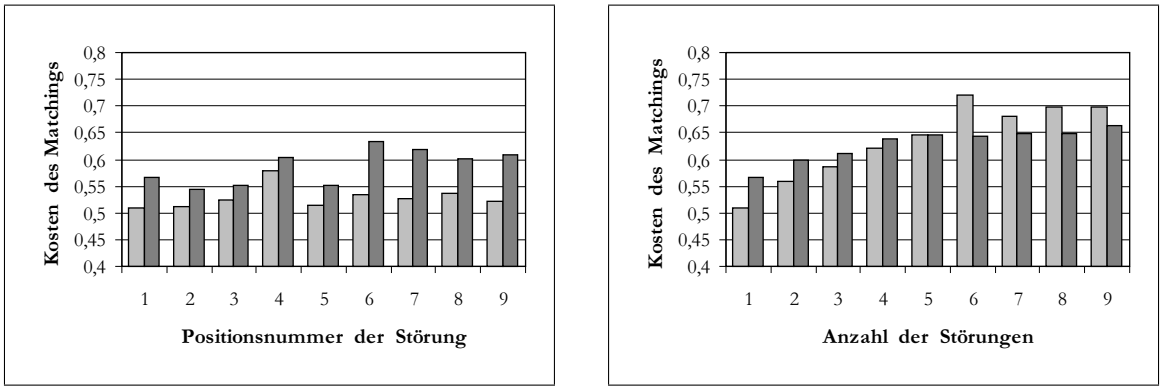
Wichtig ist die Feststellung, dass sich die Kosten für nicht genutzte Teile c^{us} erhöhen, sobald Striche im Zeilenraum auftreten, die nicht zum Wort gehören. Damit beeinflusst der Faktor $f_{\mathrm{w}}^{\mathrm{us}}$ die Ergebnisse des Tests. Er wurde bei einem Wert von 0,7 belassen – der Wert, der für diese Schrift als geeignet ermittelt wurde.

Die resultierenden Kosten des korrekten Matchings und die Differenz zum besten falschen Matching wurden erfasst und sind in Abbildung 7.13 dargestellt.



(a) Ohne Störung. (b) Eine Störung an Pos. 4. (c) Störungen an Pos. 1 bis 5.

Abbildung 7.12: Ein deutsches Beispielwort wird mit Strichen an unterschiedlichen Stellen und mit unterschiedlicher Anzahl gestört.



(a) Beispielwort wurde an unterschiedlichen Stellen von je einem Strich gestört.

(b) Beispielwort wurde von einem bis neun Strichen gestört.

Abbildung 7.13: Vergleich der Kosten zwischen der korrekten Lösung (hellgrau) und dem besten falschen Prototypen (dunkelgrau) am Beispiel des in Abbildung 7.12 dargestellten Wortkandidaten. Die künstlich zugeführten Störungen in Form von Strichen, die durch das Wort verlaufen, verhindern bis zu einer Zahl von fünf nicht die korrekte Erkennung.

7.4.2 Anpassung an neuen Schriftstil

Existieren Prototypen für eine bestimmte Schriftart, so besteht die Möglichkeit, diese an einen Schriftstil eines anderen Dokuments anzupassen, wenn die gleiche Schriftart vorliegt. Wie in Abschnitt 6.4 beschrieben, existieren mehrere Wege, diese Anpassung durchzuführen. Im Folgenden

werden die Prototypen, die mit Hilfe der Schriftproben der Wegenstedter Kirchenbücher gewonnen wurden, auf einige Monatsnamen angewandt, die aus Kirchenbüchern der St. Johannis Gemeinde in Schönebeck-Bad Salzelmen stammen. Es handelt sich dabei um lateinische Monatsnamen aus dem Ende des 18. Jahrhunderts.

Unter anderem wird die Anpassbarkeit der Prototypen mittels parametrischer Adaptation durchgeführt. Dies wird demonstriert, indem das Verhältnis zwischen der Höhe von Minuskel und Majuskel angepasst wird. Ein Schriftstil der Schönebecker Bücher zeichnet sich dadurch aus, dass er eine sehr geringe oder keine Differenz zwischen Minuskel- und Majuskelhöhe aufweist. Unglücklicherweise liefern die zur Verfügung stehenden Daten keinen kompletten Satz aller Monate mit diesen Eigenschaften. Es wurden Beispielworte von vier Monatsnamen herangezogen: März, Mai, Oktober und Dezember. Abbildung 7.14 zeigt die Schriftstile der Wegenstedter und Schönebecker Bücher anhand von Beispielen.

In den Tests wurden drei Möglichkeiten der Anpassung durchgeführt:

Manuelle Anpassung (I). In einem ersten Schritt werden Anpassungen manuell vorgenommen, die aufgrund einer anderen Schreibweise des Wortes „Mai“ bzw. „Maii“ und der individuellen Art der Schreibung des Buchstabens „t“ in „Octobre“ erforderlich waren.

Anpassung der Schriftparameter (II). Die Höhe der Majuskel wurde entsprechend den vorliegenden Daten in allen Prototypen verringert.

Automatische Primitiv Anpassung (III). Mittels eines Beispielwortes erfolgt eine Korrektur der Parameter der Primitive und ermöglicht so ein besseres Matching mit geringeren Kosten und erhöhter Robustheit.

Die geringe Zahl der Testdaten erschwert eine Analyse auf der Basis der Erkennungsrate. Daher werden die zur Bewertung der Hypothesen berechneten Kosten des Matchings zwischen Kandidaten

	Kostendifferenz (Erkennungsrate in %)			
Monat (Anzahl)	Keine Anpassung	Anpassung I	Anpassung II	Anpassung III
März (4)	**0,0700** (100)		**0,0988** (100)	**0,1351** (100)
Mai (12)	**-0,0054** (42)	**0,0530** (67)	**0,0690** (83)	**0,0971** (92)
Oktober (6)	**-0,0020** (67)	**-0,0120** (50)	**0,0283** (100)	**0,0246** (100)
Dezember (6)	**0,0294** (67)		**0,0857** (100)	**0,1079** (100)

Tabelle 7.6: Testergebnisse der Anpassung der Prototypen über drei Stufen. *Keine Anpassung* bezieht sich auf die unveränderten Prototypen, die aus den Wegenstedter Daten der Gruppe B gewonnen wurden. *Anpassung I* bedeutet das Hinzufügen oder Ersetzen von Primitiven, *Anpassung II* ist die parametrische Anpassung und *Anpassung III* bezieht sich auf die automatische Adaptation mittels Beispielwort.

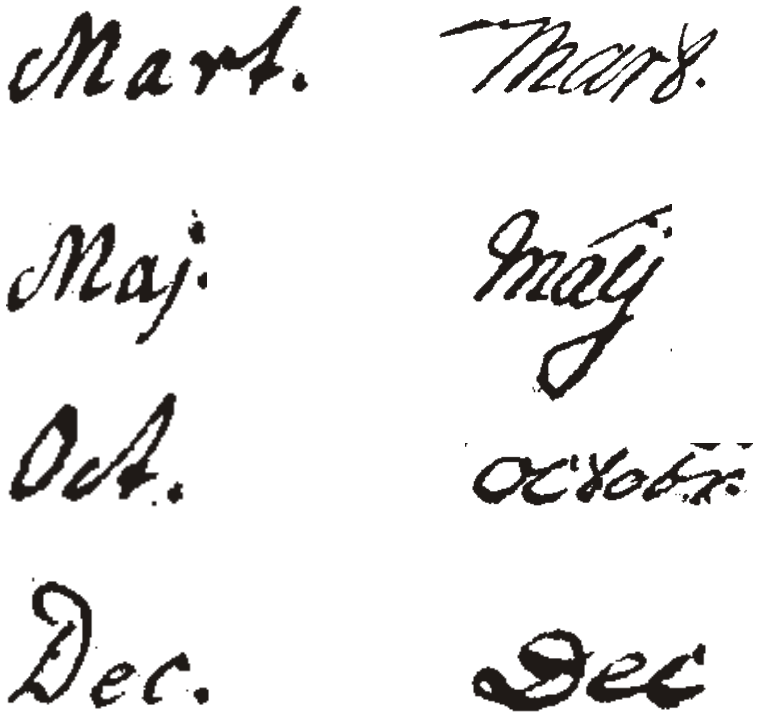


Abbildung 7.14: Vergleich zweier Schriftstile. Links die Schrift aus Wegenstedter Büchern von 1768 bis 1807 (Gruppe B), aus denen die Prototypen der lateinischen Schriftart generiert wurden. Rechts die Schrift aus Schönebecker Büchern von 1766 bis 1780 (Gruppe D), auf die die Prototypen angepasst werden.

und Prototypen betrachtet und verglichen. Eine Verbesserung des Verfahrens kann so durch die Verringerung der Kosten des entsprechenden korrekten Prototypen im Vergleich zum besten falschen Prototypen erkannt werden.

Die Testergebnisse dieser schrittweisen Anpassung sind in Tabelle 7.6 dargestellt. Tabelle 7.7 zeigt die resultierenden Prototypen.

Hinzufügen und Entfernen von Primitiven

Der Test zur Anpassung von einzelnen Primitiven auf einen neuen Schriftstil wurde für die Worte der Monate Mai und Oktober durchgeführt.

In den Schönebecker Aufzeichnungen wird der Genitiv der lateinischen Form genutzt, die auf Doppel-i endet. Der Monat Mai wurde aufgrund der geringen Wortlänge nicht abgekürzt. Das zweite i wird hierbei mit Unterlänge geschrieben [26]. Hier ist eine Anpassung der Schreibweise hilfreich, in der das zusätzliche i durch zwei Primitive dem Prototypen hinzugefügt werden.

Desweiteren wurde die Art, wie das kleine t des Oktober geschrieben wurde dem Schönebecker Stil angepasst. Der aus der Ligatur zum vorherigen Buchstaben resultierenden Aufstrich existiert in den Schönebecker Daten nicht. Statt dessen wurde der nach links oben laufende Schulterstrich des

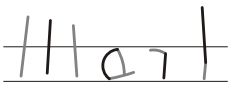
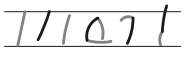
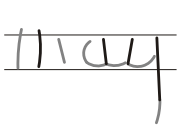
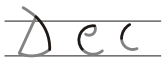
Monat	Keine Anpassung	Anpassung I	Anpassung II	Anpassung III
März				
Mai				
Oktober				
Dezember				

Tabelle 7.7: Übersicht über die Prototypen, die durch unterschiedliche Verfahren adaptiert wurden. *Keine Anpassung* stellt die Prototypen dar, wie sie aus den Daten aus Wegenstedt der Gruppe B gewonnen wurden. *Anpassung I* resultiert aus dem Hinzufügen oder Ersetzen von Primitiven. *Anpassung II* demonstriert die parametrische Anpassung. *Anpassung III* steht für die automatische Adaptation mittels Beispielwort.

kleinen „t“ in den Prototypen übernommen. Ebenso wurde die Schleife, die vom unteren Schaftende zum Schulterstrich führt, übernommen.

In Tabelle 7.6 sind die Resultate dieser Veränderungen in der Spalte „Anpassung I“ zu finden. Es fällt auf, dass sich für den Oktober die Kostendifferenz zwischen dem entsprechenden Oktober-Prototypen und dem besten falschen Prototpen durch die Veränderung verschlechtert. Der Grund dafür liegt in der Tatsache, dass die Prototypen so angepasst wurden, dass sie einem Schriftstil entsprechen, der einen deutlichen Unterschied zwischen Minuskel und Majuskel aufweisen – so wie dies bei den Wegenstedter Daten der Fall ist. Im Bezug zu den Schönebecker Daten, auf die die Prototypen angepasst werden sollen, passen die Primitive des zu hohen „t“ noch schlechter als zuvor. Dieser Schritt war erforderlich, um die Wirkungsweise der anschließenden Anpassung der Schriftparameter zu zeigen.

Für den Monat Mai betraf die Anpassung keinen Abschnitt mit Oberlänge. Die Ergebnisse zeigen, wie sich die Kostendifferenz von -0,0054 auf +0,053 verbessert. In Folge dessen steigt die Erkennungsrate von 42 % auf 67 %.

Anpassung der Schriftparameter

Zur Demonstration der parametrischen Anpassbarkeit wurde die Größe der Oberlängenabschnitte korrigiert. Bei allen Prototypen wurden die Höhe der Oberlängen auf 57 % reduziert. D. h. die Primitive des Abschnitts wurden vertikal skaliert, indem die Punkte, die sie definieren (S, H, E) in Abhängigkeit vom Abstand zur Basislinie vertikal verschoben wurden.

Die Veränderung der Größe der Majuskel in Relation zu den Minuskeln dient als Beispiel für die Anpassbarkeit der Prototypen über Parameter der Schrift. Normiert zu den Kosten des entsprechenden Prototypen verbesserte sich die Kostendifferenz im Mittel von 8,3 % auf 15,2 %. Die Erkennungsrate stiegt von 71 % auf 96 %. Diese Ergebnisse zeigen, dass sich die Qualität der Prototypen für den gegebene Schriftstil verbessert.

So wie hier die Primitive der Oberlängen ausgewählt und entsprechend angepasst wurden, kann dies mit jeder anderen Gruppe von Primitiven geschehen – je nach dem, welche Merkmale eine Schrift aufweist.

Anpassung bestehender Primitive

Die nach der zweiten Stufe angepassten Prototypen werden durch die Nutzung eines Beispielwortes für die Schönebecker Schrift weiter verbessert.

In Tabelle 7.6 unter „Anpassung III“ sind die Resultate dieser Optimierung dargestellt. Ebenso präsentiert Tabelle 7.7 die resultierenden Prototypen in dieser Spalte.

Um eine Mittelung der Primitiv-Parameter zu erhalten wurde ein Anpassungsfaktor von 0,5 gewählt. An dieser Stelle wurden keine weiteren Durchläufe mit unterschiedlichen Faktoren durchgeführt, da dieser Wert sehr stark von den gegebenen Daten und deren Varianz abhängt und keine generellen Erkenntnisse gewonnen werden können.

Es zeigt sich in drei von vier Anpassungen eine Verbesserung der Kostendifferenz. Lediglich beim Monat Oktober kommt es zu einer leichten Verschlechterung des Wertes. Daraus kann geschlossen werden, dass durch die vorherigen Anpassungen I und II ein Prototyp vorliegt, dessen Primitive die optimalen Parameter besitzen, um die Kandidaten der Schönebecker Schrift zu repräsentieren. Eine Anpassung durch einen Kandidaten führt zu einer Spezialisierung, durch die der Prototyp seinen generellen Charakter verliert. Für die anderen Kandidaten bedeutet dies eine Erhöhung der Matching Kosten.

Dass es zu einer Verbesserung des Matchings zwischen dem Prototypen und dem Kandidaten kommt, der zur Optimierung herangezogen wird, liegt auf der Hand. Um zu zeigen, dass die Anpassung auch für die anderen Kandidaten des gleichen Schriftstils eine Verbesserung bedeutet, wurde ein weiterer

Testlauf ohne die Kandidaten durchgeführt, die zur Optimierung genutzt wurden. Die Ergebnisse, die in Tabelle 7.8 aufgeführt sind, bestätigen die Ergebnisse aus Tabelle 7.6.

7.4.3 Fehlerquellen

Die in deutscher Schrift verfassten Daten stammen aus dem Jahren 1807 bis 1816. Hier zeigt sich die Schwierigkeit der Unterscheidung zwischen Monaten, die ein ähnliches Erscheinungsbild haben. So gab es Schwierigkeiten, den Monat Januar zu erkennen, da er der Schreibweise des Juni innerhalb der ersten drei bis vier Buchstaben stark ähnelt. Es wurden lediglich 20 % der Januar Kandidaten korrekt erkannt. Eine Verbesserung konnte durch Ergänzung des zunächst nur aus den ersten drei Buchstaben bestehenden Prototypen des Januar und des Juni erreicht werden. Diese Anpassung ging einher mit der Erhöhung der Kosten für nicht zugewiesene Striche des Kandidaten. Dadurch erhöhten sich die Gesamtkosten des Matchings des Juni Prototypen mit einem ausgeschriebenen Januar Kandidaten. Erreicht wurde dadurch eine Erkennungsrate von 87 % für den Monat Januar. Nimmt man den Zweitplatzierten hinzu, wurden alle Kandidaten richtig erkannt. Der Juni wurde unverändert in 75 % (Erstplatzierte) bzw. 88 % (Erst- und Zweitplatzierte) der Fälle erkannt.

7.5 Erkennung des kompletten Datums

Das in Kapitel 4 beschriebene Verfahren zur Segmentierung der Worte und Ziffern liefert mehrere Hypothesen von Aufteilungen des Datums. Die Zahl der Hypothesen, die in Betracht gezogen werden, hängt von der Art und der Qualität der gegebenen Schrift ab sowie von der zur Verfügung stehenden Rechenleistung. Für jedes vermutete Objekt (Ziffer, Monatsname) wird die entsprechende Erkennung durchgeführt. Das Ergebnis ist eine Vielzahl von möglichen Lösungen. Für jedes Objekt in jeder Hypothese gibt es eine Liste mit möglichen Interpretationen. Durch das Kombinieren dieser Objekt-Hypothesen werden komplette Vorschläge erzeugt, deren Bewertungen sich aus den

	Kostendifferenz (Erkennungsrate in %)	
Monat (Anzahl)	Anpassung II	Anpassung III
März (4)	**0,0958** (100)	**0,1283** (100)
Mai (12)	**0,0608** (82)	**0,0971** (91)
Oktober (6)	**0,0284** (100)	**0,0267** (100)
Dezember (6)	**0,0710** (100)	**0,0916** (100)

Tabelle 7.8: Testergebnisse der Anpassung III: Anpassung bestehender Primitive. Bei diesem Durchlauf wurden die Kandidaten, die zur Anpassung genutzt wurden nicht betrachtet.

Kosten der einzelnen Objekt-Hypothesen ergeben. Die Kosten dieser Hypothesen werden dabei gemittelt. Tests, bei denen Ziffern und Wörter mit unterschiedlichen Gewichten in die Bewertung des kompletten Datums eingingen, ergaben keine Verbesserung.

Ein Schwierigkeit stellt das Vergleichen von Hypothesen unterschiedlicher Formen dar. Auch wenn keine Erkennung der Artefakte erfolgt, muss doch die potentielle Existenz eines Artefakts in die Berechnung mit einbezogen werden. Anderenfalls werden bei einem vorliegenden Datum der Form *C-C-M* die Hypothesen der Form *C-A-M* besser bewertet. Um die Vergleichbarkeit zu verbessern, wurde ein konstanter Kostenwert für ein potentielles Artefakt eingeführt. Ist dieser Wert im Verhältnis zu den Kosten der Ziffern zu klein, kommt es häufiger zu der oben beschriebenen falschen Bewertung bei der Form *C-C-M*. Liegt der Wert des Artefakts zu hoch, wird umgekehrt die Form *C-A-M* als *C-C-M* erkannt, da das Artefakt als Ziffer interpretiert wird. Versuche mit allen Schriften ergab ein gutes Resultat bei einem Wert von $\mathrm{c}_{\mathrm{artefact}} = 0{,}4$.

Nachdem eine Bewertung und damit eine Reihenfolge der Hypothesen über das vorliegende Datum vorliegt, werden durch die Anwendung von Regeln nicht mögliche Datumsangaben entfernt. Stehen weitere Informationen zur Verfügung, können weitere Lösungen ausgeschlossen oder als unwahrscheinlich eingestuft werden. Dies ist beispielsweise möglich, wenn die zu verarbeitenden Daten chronologisch vorliegen.

7.5.1 Abschätzen der zu erreichenden Erkennungsrate

Anhand der Ergebnisse der Teilprozesse kann abgeschätzt werden, mit welcher Wahrscheinlichkeit die korrekte Hypothese des Datums die beste Bewertung erhält. Anhand dieses Schätzwertes können Aussagen über die Qualität des Prozesses zur Kombination der einzelnen Teilergebnisse gemacht werden.

Hierfür ist es erforderlich, eine getrennte Betrachtung der Stilgruppen durchzuführen. Für die drei gegebenen Stilgruppen wurden Tests zur Bildung der Hypothesen der Wortgrenzen sowie zur Ziffernerkennung durchgeführt. In Verbindung mit den Ergebnissen der Erkennung der Monatsnamen lässt sich eine obere Schranke der zu erwartenden Gesamterkennungsrate abschätzen.

Die Erkennungsrate des gesamten Datums hängt von folgenden Größen ab:

Wahrscheinlichkeit der korrekten Hypothese der Wortgrenzpositionen *p(GH)*. In Abhängigkeit von der Schriftqualität einerseits und der Rechenzeit andererseits, werden aus den Ergebnislisten der Hypothesen der Wortgrenzen für die weitere Erkennung die besten zwei, drei oder vier betrachtet. Im Zusammenhang mit den in Tabelle 7.1 auf Seite 82 dargestellten Ergebnissen ergibt sich daraus ein Wahrscheinlichkeitswert, ob die korrekte Hypothese unter den ausgewählten ist oder nicht.

Erkennungsrate der Ziffern *p(Ziffer)*. Kann eine Ziffer nicht korrekt erkannt werden, ist die komplette Erkennung des Datums falsch. Diese Abhängigkeit quadriert sich, wenn sich zwei Ziffern im Datum befinden und erkannt werden müssen.

Erkennungsrate der Monatsnamen *p(Monat)*. So wie die Erkennung der Ziffern beeinflusst auch die Erkennung der Monatsnamen das Gesamtergebnis.

Es ergibt sich folgender Zusammenhang:

$$p\,(C\text{-}M) = p\,(C\text{-}A\text{-}M) = p\,(GH) \cdot p\,(\mathit{Ziffer}) \cdot p\,(\mathit{Monat}) \tag{7.1}$$

$$p\,(C\text{-}C\text{-}M) = p\,(C\text{-}C\text{-}A\text{-}M) = p\,(GH) \cdot p\,(\mathit{Ziffer})^2 \cdot p\,(\mathit{Monat}) \tag{7.2}$$

Für die Testläufe wurden aus den Ergebnislisten der Hypothesen der Wortgrenzen jeweils die zwei besten Hypothesen für die weitere Verarbeitung herangezogen. Damit sind es acht Hypothesen, die im weiteren Verlauf betrachtet werden. Damit werden im Mittel 86,4 % der korrekten Hypothesen der Wortgrenzen erfasst (siehe Tabelle 7.1 auf Seite 82).

	Gruppe A Latein 1714–1730	Gruppe B Latein 1768–1807	Gruppe C Deutsch 1807–1816	Gesamt
Korrekte Hypothesen der Wortgrenzen in %	81,3	98,0	79,9	86,4
Erkennungsrate der Ziffern in %	65,7	89,6	81,4	79,9
Erkennungsrate der Monatsnamen in %	78,9	91,2	83,4	84,7
Erkennungsrate des Datums mit einer Ziffer in %	42,1	80,1	54,3	58,5
Erkennungsrate des Datums mit zwei Ziffern in %	27,7	71,8	44,2	46,7
Erkennungsrate des Datums in %	**31,9**	**73,8**	**47,1**	**50,0**

Tabelle 7.9: Abschätzen der erreichbaren Erkennungsrate des gesamten Datums.

Position in Ergebnisliste	Anzahl der korrekten Hypothesen an dieser Position				Rate in %	Akkumulierte Rate in %
	Latein 1714–1730	Latein 1768–1807	Deutsch 1808–1815	Gesamt		
1	32	98	71	201	46,2	**46,2**
2	10	13	26	49	11,3	**57,5**
3	10	7	13	30	6,9	**64,4**
4	5	3	6	14	3,2	**67,6**
5	7	2	4	13	3,0	**70,6**
6	9	1	2	12	2,8	**73,3**
7	3	0	4	7	1,6	**74,9**
8	1	0	1	2	0,5	**75,4**
9	0	1	0	1	0,2	**75,6**
> 9	46	23	37	106	24,4	**100,0**
Gesamt	123	148	164	435		

Tabelle 7.10: Ergebnisse der kompletten Erkennung mit 4 x 2 betrachteten Hypothesen der Wortgrenzen.

7.5.2 Ergebnisse der kompletten Erkennung

Die Auswertungen zeigen, dass das Zusammenfügen der Teilergebnisse funktioniert. Es werden annähernd die in Tabelle 7.9 dargestellten Schätzwerte erreicht. Betrachtet man die beiden besten Hypothesen, so liegt der Wert sogar über der Abschätzung. Ein Grund hierfür dürfte auch die nachträgliche Bereinigung der Ergebnisliste von Hypothesen darstellen, die kein gültiges Datum bilden. Dass dennoch mit dem Erstplatzierten die optimale Erkennungsrate nicht erreicht wird, hat mehrere Ursachen. Ein wesentlicher Grund sind Bestandteile des Datums, die nicht durch eine Erkennung überprüft werden („ten“ oder „te“). Dadurch kommt es einerseits zu Hypothesen, bei denen dieses Artefakt als eine Ziffer interpretiert wird und andererseits zu Hypothesen, bei denen die zweite Ziffer oder Teile des Monatsnamen als Artefakt vermutet werden. In beiden Fällen kann die Bewertung dieser Hypothesen besser ausfallen als die korrekte Lösung.

Kapitel 8

Fazit und Ausblick

Nur wer die Vergangenheit kennt,
hat eine Zukunft.

Wilhelm von Humboldt (1767-1835)

Auf dem Gebiet der Schrifterkennung wurden in den letzten Jahren viele Fortschritte erzielt. Aber es musste auch erkannt werden, dass Handschrift viel schwieriger zu verarbeiten ist als gedruckter Text. Dies ist einer von mehreren Gründen, weshalb im Bezug auf die Verarbeitung von historischen Dokumenten verhältnismäßig wenig Ansätze existieren. Im Hinblick auf die immer weiter fortschreitende Digitalisierung von ganzen Bibliotheken alter Dokumente liegt es nahe, diese Daten weiterzuverarbeiten und nicht nur als hochauflösende Farbbilder ohne jegliche strukturelle und inhaltliche Information abzuspeichern.

8.1 Fazit

Mit der hier verfassten Arbeit wird ein wichtiger Beitrag zur Verarbeitung von Schrift in historischen Dokumenten geleistet. Es wurden Techniken sowohl zur Segmentierung als auch zur Erkennung von Schrift dieser besonderen Form entwickelt und getestet. Dabei wurde besonderes Augenmerk auf die Möglichkeiten der Nutzung gelegt, ohne zuvor ein aufwendiges Training durchführen zu müssen. Dies ist wichtig, da gerade das Fehlen eines geeigneten Trainingsdatensatzes die Verwendung herkömmlicher statistischer Verfahren in Frage stellt. Sowohl die Segmentierung als auch die Erkennung von Ziffern und Wörtern bieten die Möglichkeiten einer Anpassung an eine vorliegende Schrift mittels Merkmalen wie Schriftgröße und Schriftform.

Aufgrund der Tatsache, dass sich anhand der geometrischen Konstellationen der Textobjekte keine eindeutigen Aussagen über die Position der Wortgrenzen machen lassen, werden mehrere Hypothesen über potentielle Lösungen betrachtet. Bei jedem weiteren Verarbeitungsschritt werden die neu

gewonnenen Informationen genutzt, um diese Hypothesen zu bewerten und somit unwahrscheinliche von den wahrscheinlichen zu trennen.

Anhand der Ziffernerkennung wurde eine Metrik entwickelt, die einen Zusammenhang zwischen der *Abweichung* zweier Striche und dem Begriff der *Ähnlichkeit* herstellt. Es wird demonstriert, dass anhand von Prototypen Striche als Merkmal für ein Klassifizieren von Ziffern und Wörtern auch in alten Dokumenten möglich ist. Darüber hinaus zeigt die Arbeit, dass die Nutzung eines strukturellen Verfahrens zur Erkennung von Ziffern und Wörtern entscheidende Vorteile gegenüber dem statistischen Ansatz bietet. Das Verfahren kann eingesetzt werden, ohne zuvor ein Training mit tausenden von Testdaten durchgeführt zu haben. Liegt eine Schrift in einem bestimmten Schriftstil vor, können bestehende Prototypen auf mehrere Arten angepasst werden. Dabei besteht die Möglichkeit, paläografisches Wissen zur Anpassung zu nutzen. Es konnte weiterhin gezeigt werden, dass das Erkennungsverfahren auch in schwiegen Situationen, in denen das Schriftbild durch Artefakte wie Strichen aus benachbarten Zeilen gestört ist, Worte korrekt erkennt.

Weitere Erkenntnisse dieser Arbeit betreffen die besonderen Merkmale alter Schriften. Das Vorhandensein störender Striche findet in anderen Arbeiten kaum Beachtung. Weiterhin kann nicht vorausgesetzt werden, dass ein zu erkennender Wortkandidat an den entsprechenden Positionen die zu erwartenden Ausprägungen einer Ober- oder Unterlänge aufweist. Diese Besonderheiten wurden in dieser Arbeit erkannt und berücksichtigt.

Durch das Erkennen des Datums in Kirchenbucheinträgen wurde zum Einen die Funktionsweise des Verfahrens praktisch demonstriert, zum Anderen gehört gerade das Datum in diesen Aufzeichnungen zu den entscheidenden Informationen, die für einen Interessenten von Bedeutung sind. Diese Arbeit liefert die Möglichkeit, historische Dokumente mit ihren besonderen Eigenschaften zu verarbeiten. Neben Kirchenbücher können dies beispielsweise Kopialbücher oder Ratsprotokolle sein, denn auch hier liegen wichtige Informationen vor, die meist in einer bestimmten Form und Anordnung auf das Papier gebracht wurden.

8.2 Erweiterungen und Verbesserungen

Die Arbeit zeigt einen neuen Weg, Informationen aus historischen Dokumenten mit Hilfe des Rechners zu gewinnen. Da hier die Machbarkeit des gewählten Ansatzes im Vordergrund stand, konnten nicht alle Aspekte, die eine Maximierung der Erkennungsrate und Effizienz zum Ziel haben, erschöpfend diskutiert werden. Eine Vielzahl von Veränderungen und Erweiterungen sind denkbar. In erster Linie besteht das Bestreben, die Erkennungsleistung weiter zu verbessern. Ein weiteres Anliegen ist die Erhöhung der Effizienz. Muss zu lange auf ein Ergebnis gewartet werden, verringert sich der praktische Nutzen der Anwendung. Darüber hinaus existieren Möglichkeiten der Erweiterung, die diesen Nutzwert erhöhen können, da sie die Bedienung erleichtern. Diese weiterführenden Überlegungen werden im Folgenden kurz vorgestellt.

Automatische Selektion des Datums

Die Position des Datums muss vom Nutzer manuell markiert werden, damit eine Erkennung erfolgen kann. Es wäre eine Erweiterung wünschenswert, die diese Interaktion nicht notwendig macht.

Das Datum eines Ereignisses befindet sich meist am Anfang des Eintrags[1] oder am Ende[2]. Nachdem die Zeilen erfolgreich segmentiert wurden, ist es denkbar, dass ein entsprechendes Verfahren die Position des Datums automatisch ermittelt und eine Segmentierung durchführt. Das Ergebnis wäre eine Vereinfachung und Beschleunigung des Erkennungsprozesses.

Besseres Matching durch variablen Bezugspunkt

Diese Überlegung bezieht sich auf die Berechnung der Translationskosten c^{tr} zwischen zwei Primitiven, wie sie in Abschnitt 5.5.2 beschrieben wurde (siehe S. 55 ff.). Die Translation wird auf der Basis des Start- oder des Endpunktes eines Primitivs berechnet. Diese Punkte wurden gewählt, da sie eindeutig und leicht bestimmt werden können.

Sind zwei Primitive so angeordnet, dass die Stelle der größten Nähe nicht an einem der Endpunkte liegt, so führen leichte Abweichungen der Form, wie z. B. eine leichte Schriftneigung, in einem größeren Maße zu Translationskosten, als dies bei Primitiven der Fall ist, deren Bezugspunkt mit dem Punkt der größten Nähe übereinstimmt. Die Abweichungen von Translation und Rotation können während eines Matchings besser durch einen frei wählbaren Bezugspunkt auf dem Primitiv erfasst werden.

Verringern des Rechenaufwands

Je effizienter ein Verfahren arbeitet, umso größer ist das Potential, die Erkennungsrate zu erhöhen. Charakteristisch für strukturelle Verfahren ist der verhältnismäßig hohe Rechenaufwand. Sehr viele Kombinationen von Merkmalen werden gebildet und getestet. Bei den längsten Wörtern und einer hohen Zahl von Abschnitten, nach denen im Wort gesucht wird, kann auch auf einem schnellen Rechner die Worterkennung mehrere Sekunden in Anspruch nehmen.

Es besteht ein großes Potential zur Erhöhung der Geschwindigkeit des Verfahrens. Die Daten setzen sich aus Wörtern zusammen, die vollständig und frei von Artefakten sind und Wörtern, bei denen nicht alle Merkmale erkannt werden können. In jeder Phase der Verarbeitung – von der Approximation bis zum Matching – sind die Toleranzen so gewählt, dass auch ungünstige und gestörte Kandidaten erkannt werden.

[1] Bei den zur Verfügung stehenden Daten gilt dies für den Zeitraum Ende 18. Anfang 19. Jahrhundert.
[2] Die Daten vom Anfang des 18. Jahrhunderts zeigen dieses Merkmal.

So wie die Geschwindigkeit des menschlichen Lesens von der Qualität der Schrift abhängt [83], kann auch der hier vorgestellte strukturelle Worterkenner so erweitert werden, dass bei sauber geschriebener Schrift sehr viel weniger Zeit zur Erkennung benötigt wird. Es ist denkbar, das Verfahren so zu erweitern, dass mehrere Durchläufe zur Erkennung eines Kandidaten möglich sind. Was zunächst wie eine Erhöhung des Rechenaufwands erscheint, sollte zu einer Verringerung führen, indem in der ersten Stufe erwartet wird, dass ein ungestörter und vollständiger Kandidat vorliegt. Die Toleranzen werden somit sehr klein gewählt, wodurch die Rechenzeit auf ein Bruchteil reduziert werden kann. Liegen die dadurch entstehenden Kosten über einer gewissen Schwelle, wird der Vorgang mit erhöhten Toleranzen wiederholt. Das System reagiert dynamisch auf die Qualität der vorliegenden Schrift.

Reduzierung des Lexikons

Bevor die Erkennung eines Kandidaten durchgeführt wird, ist es möglich, die Zahl der in Frage kommenden Prototypen zu reduzieren. Unterschiedliche Verfahren wurden beschrieben, um vor allem in Systemen mit großen Lexika den Rechenaufwand zu verringern [53, 104].

Eine einfache Möglichkeit der Reduzierung des Lexikons kann durch ein Betrachten der Wortlänge erfolgen [52]. Momentan überprüft ein Prototyp lediglich das Vorhandensein bestimmter Merkmale. Eine Verbesserung der Erkennungsleistung kann erzielt werden, wenn die zu erwartende Breite eines Wortes mit im Prototypen gespeichert wird. Auch wenn Monatsnamen abgekürzt auftreten, so gibt es für ein Wort eine maximal zulässige Länge. Ist die Breite eines Kandidaten deutlich größer als die von einem Prototypen erwartete, so ist die Wahrscheinlichkeit gering, dass es sich um dieses Wort handelt. Die Bewertung nicht genutzter Striche, wie sie in dieser Arbeit umgesetzt wurde, kann diese Bewertung nicht ersetzen.

Zur Verdeutlichung ein Beispiel: Es liegt ein zu erkennender ausgeschriebener Kandidat des Wortes „Januar“ vor. Durch kleine Abweichungen oder Artefakte kann sehr leicht eine besseres Matching des Juni-Prototypen erfolgen als des korrekten Januar-Prototypen. Wird im Vorfeld eine Bewertung der Wortbreite vorgenommen, können mehrere Prototypen aus der Erkennung ausgeschlossen werden. Sämtliche kurze Worte des Lexikons kommen nicht in Frage. Dies erhöht außerdem die Verarbeitungsgeschwindigkeit.

Reduzierung störender Striche durch Strichsegmentierung

Obwohl das entwickelte Verfahren eine Robustheit gegenüber störenden Strichen aufweist, besteht das Potential, die Performance des Erkenners zu erhöhen, indem eine Identifizierung sich überkreuzender Striche erfolgt. Striche, die aus benachbarten Zeilen in den Zeilenraum des zu erkennenden Textes ragen, werden bereits erfolgreich der jeweiligen Zeile zugeordnet und beeinflussen somit nicht

den Erkennungsprozess negativ. Kommt es jedoch zu Überschneidungen von Federzügen benachbarter Zeilen, so erschwert dies die Erkennung. Erste Ansätze zur Zuordnung der Striche auch in Fällen der Strichüberkreuzung wurden bereits in [17] umgesetzt. Weiterführende Untersuchungen können die Robustheit des gesamten Systems weiter erhöhen, indem Techniken zur Strichsegmentierung, wie sie in [68] erläutert werden, auf das Problem der sich berührenden Zeilen angewendet werden. Da der Ursprung eines Striches einer Unterlänge meist bekannt ist, können so die Teile dieses Striches auch über Kreuzungen hinweg identifizieren und gegebenenfalls entfernt werden.

Segmentbasiertes Prototyping

Das hier vorgestellte strukturelle Verfahren folgt in erster Linie dem holistischen Ansatz einer Worterkennung. Beim automatischen Prototyping werden komplette Wörter analysiert und Merkmale extrahiert.

Im Hinblick auf die Möglichkeiten einer Anpassung der Prototypen an einen neuen Schriftstil ist es wünschenswert, eine Anpassung einzelner Buchstaben zu ermöglichen. Paläografische Arbeiten beschreiben vor allem die Merkmale einzelner Buchstaben, wie diese Zitate verdeutlicht [61]:

> *„Ab dem 17. Jahrhundert neigt sich die Schriftachse mehr nach rechts, bis sie im 19. Jahrhundert schon ziemlich schief lag. Das Minuskelalphabet stabilisierte sich. Das e wurde weitgehend dem n angeglichen und nur beide Schäfte enger aneinandergerückt. Die Spitzform des c ähnelt einem i ohne Punkt; g und h sind immer öfters in Schleifenform vertreten; die Bäuche von a, o, q bleiben offen.“*

Solche Informationen sind nutzbar, wenn Prototypen für Buchstaben erzeugt werden. Zur Worterkennung werden diese Prototypen kombiniert. Besonders Schriften ohne Ligaturen sind dafür geeignet.

Ob mit diesem Ansatz auch deutsche Schriften, wie die hier zur Verfügung stehenden, verarbeitet werden können, muss untersucht werden, da bei dieser Schriftart das Aussehen eines Zeichens stark von dem vorhergehenden und nachfolgenden Zeichen abhängt.

Automatisches Erfassen der Schriftparameter

In der aktuellen Umsetzung des Erkennungssystems ist es erforderlich, neben der Auswahl der dem Schriftstil entsprechenden Prototypen die rudimentären Parameter der Schrifthöhe und -breite anzugeben. Es existieren bereits Verfahren, die aufgrund des Schriftbildes diese Werte extrahieren [10]. Sollte es gelingen, diese Ansätze auf Schriften der Kirchenbücher anzuwenden, reduziert sich die Menge der Informationen, die ein Nutzer vorab anzugeben hat.

Werkzeuge zur Erzeugung und Anpassung der Prototypen

In dieser Arbeit steht die Verwendung eines strukturellen Erkennnungsverfahrens im Mittelpunkt. Um aus den hier gewonnenen Erkenntnissen ein gut nutzbares Werkzeug für einen Historiker, Genealogen oder Paläografen zu entwickeln, ist es sinnvoll, die Erzeugung und Manipulation der Prototypen zu erleichtern.

Die Daten der Prototypen sind in einer Datei mit allen erforderlichen Positionswerten und Merkmalen der Abschnitte und Primitive gespeichert, die manuell editiert werden muss, um Änderungen an den Prototypen vorzunehmen. Diese Arbeit kann erheblich vereinfacht werden, wenn eine grafische Oberfläche zur Verfügung steht, in der die Primitive der Prototypen angezeigt, ausgewählt, manipuliert, erzeugt und gelöscht werden können. In Kombination mit dem hinterlegten Bild eines Wortkandidaten kann so sehr leicht ein neuer Prototyp erzeugt oder ein bestehender angepasst werden.

Verbesserung der Bewertung der Datumshypothesen

Die Ergebnisse der Erkennung des kompletten Datums zeigen das erfolgreiche Zusammenspiel der einzelnen Module. Weitere Untersuchungen zur Bewertung einer Datumshypothese auf der Basis der Teilergebnisse können die Resultate weiter verbessern.

8.3 Ausblick

Die präsentierten Ergebnisse zeigen das erfolgreiche Arbeiten des strukturellen Ansatzes zur Erkennung von Handschriften in schwierigen, gestörten Situationen. Im Hinblick auf die praktische Anwendbarkeit muss jedoch festgestellt werden, dass der momentane Stand einen effektiven praktischen Einsatz zur Transkription von historischen Texten noch nicht ermöglicht.

Der strukturelle Ansatz zur Erkennung historischer Texte ist ein richtiger Schritt und ein weiteres Element auf dem Weg zur automatischen Transkription. Wird an diesem Punkt weitere Entwicklungsarbeit geleistet, kann auch unter der Berücksichtigung der oben angeführten Vorschläge in naher Zukunft ein System entstehen, das von Historikern und Genealogen, aber auch von Paläografen für ihre praktische Arbeit genutzt werden kann.

Zunächst wird es möglich sein, Dokumente in tabellarischer Formen zu verarbeiten, da hier keine Segmentierung einer Zeile in Worte notwendig ist. Das Erkennungssystem wird in der Lage sein, dem Nutzer Vorschläge über den Inhalt der einzelnen Tabellenfelder zu liefern und somit die Transkription dieser Dokumente deutlich erleichtern. Später sollten auch solche Dokumente, wie sie hier betrachtet wurden, komplett verarbeitet werden können. Liegen diese so gewonnenen Informationen erst einmal in elektronischer Form vor, können sie in Datenbanken weiterverarbeitet und auf

vielfältige Weise Interessenten zugänglich gemacht werden. Dadurch wird es möglich sein, die Informationen der in großen Zahlen in Archiven lagernden Dokumente zu erfassen und vor dem Verlust durch Zerfall zu retten.

Im Hinblick auf den paläografischen Aspekt ermöglicht der hier vorgestellte strukturelle Ansatz die Nutzung des Wissens über Schriften und ihre Entwicklung. Existiert erst einmal eine Datenbank, die die Informationen darüber enthält, wo und wann wie geschrieben wurde, in welcher Beziehung die einzelnen Schriftformen zueinander stehen und welcher Veränderung diese Schriftformen unterliegen, so ergeben sich zwei Szenarien: *(1)* Soll der Text eines alten Dokuments erkannt werden, so kann das Erkennungssystem durch die Angabe der Dokumentdaten wie Zeit, Ort und Form die passenden Prototypen auswählen bzw. entsprechende Anpassungen durchführen und somit den Text des Dokuments erfolgreich verarbeiten. *(2)* Umgekehrt ist es denkbar, dass das System für ein vorliegendes Dokument anhand der Schriftform Hypothesen über Ort und Zeit aufstellt, die einem Historiker die Bewertung und Einordnung des Dokuments erleichtern.

Auch wenn diese Vorstellungen noch als „Zukunftsmusik“ betitelt werden können, so ist es doch nicht aussichtslos, die Forschungsarbeit in diese Richtung fortzuführen. Die präsentierten Ergebnisse zeigen, dass das informationstechnische Verarbeiten alter Dokumente möglich ist und großes Potential in sich birgt. Und so ist es nicht abwegig, davon zu sprechen, dass es irgendwann Datenbanken geben wird, gefüllt mit Informationen der Vergangenheit, die durch teilweise automatische Transkription alter Dokumente gewonnen wurden.

Literaturverzeichnis

[1] Aas, K., L. Eikvil und T. Andersen: *Text Recognition from Grey Level Images Using Hidden Markov Models*. In: *Computer Analysis of Images and Patterns*, Seiten 577–580, Prague, 1995.

[2] Bozinovic, R. M. und S. N. Srihari: *Off-Line Cursive Word Recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 11(1):68–83, 1989.

[3] Brakensiek, A., A. Kosmala und G. Rigoll: *Writer Adaptation for On-Line Handwriting Recognition*. In: *Pattern Recognition, 23rd DAGM Symposium*, Seiten 32–37, Munich, Germany, September 2001. Springer-Verlag.

[4] Brakensiek, A. und G. Rigoll: *Handwritten Address Recognition Using Hidden Markov Models*. In: *Reading and Learning: Adaptive Content Recognition*, Nummer 2956 in *Lecture Notes in Computer Science*, Seiten 103–122. Springer-Verlag, 2004.

[5] Breuel, T. M.: *Segmentation of Handprinted Letter Strings Using a Dynamic Programming Algorithm*. In: *Sixth International Conference on Document Analysis and Recognition – ICDAR 2001*, Seiten 821–826. IEEE Computer Society, September 2001.

[6] Cai, J. und Z.-Q. Liu: *Off-line Unconstrained Handwritten Word Recogniton*. In: *Australian and New Zealand Conference on Intelligent Information Systems*, Seiten 199–202, Adelaide, SA, Australia, 1996.

[7] Cha, S.-H., Y.-C. Shin und S. N. Srihari: *Approximate Stroke Sequence String Matching Algorithm for Character Recognition and Analysis*. In: *Fifth International Conference on Document Analysis and Recognition*, Seiten 53–56, Bangalore, India, 1999.

[8] Correia, S., J. Carvalho und R. Sabourin: *On the Performance of Wavelets for Handwritten Numerals Recognition*. In: *16th International Conference on Pattern Recognition*, Band 3, Seiten 30127–30130, Quebec, Canada, August 2002.

[9] Côté, M., E. Lecolinet, M. Cheriet und C.Y. Suen: *Building a Perception Based Model for Reading Cursive Script*. In: *Third International Conference on Document Analysis and Recognition*, Seiten 898–901, Montréal, Canada, August 1995.

[10] Crettez, J.-P.: *A Set of Handwriting Families: Style Recognition*. In: *Third International Conference on Document Analysis and Recognition*, Seiten 489–494, Montreal, Canada, August 1995.

[11] Cun, Y. L., J. S. Denker, D. Henderson, R. E. Howard, W. Hubband und L. D. Jackel: *Backpropagation Applied to Handwritten Zip Code Recognition*. In: *Neuronal Computation*, Band 1, Seiten 541–551, 1989.

[12] Doermann, D., J. Liang und H. Li: *Progress in Camera-Based Document Image Analysis*. In: *Seventh International Conference on Document Analysis and Recognition*, Seiten 606–616, Edinburgh, UK, August 2003.

[13] El-Yacoubi, A., M. Gilloux und J.-M. Bertille: *A Statistical Approach for Phrase Location and Recognition within a Text Line: An Application to Street Name Recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(2):172–187, Februar 2002.

[14] El-Yacoubi, A., M. Gilloux, R. Sabourin und C. Y. Suen: *An HMM-Based Approach for Off-Line Unconstrained Handwritten Word Modeling and Recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(8):752–760, August 1999.

[15] Engel, Edna: *The Analysis of the Letter – a New Palaeographical Method*. In: Rück, Peter (Herausgeber): *Methoden der Schriftbeschreibung*, Band 4 der Reihe *Historische Hilfswissenschaften*, Seiten 43–50. Jan Thorbecke Verlag, 1999.

[16] Favata, J. T.: *General Word Recognition Using Approximate Segment-String Matching*. In: *Fourth International Conference on Document Analysis and Recognition*, Seiten 92–96, Ulm, Germany, 1997.

[17] Feldbach, M.: *Generierung einer semantischen Repräsentation aus Abbildungen handschriftlicher Kirchenbuchaufzeichnungen*. Diplomarbeit, Otto-von-Guericke Universität, Magdeburg, 2000.

[18] Feldbach, M. und K. D. Tönnies: *Line Detection and Segmentation in Historical Church Registers*. In: *Sixth International Conference on Document Analysis and Recognition*, Seiten 743–747, Seattle, USA, September 2001. IEEE Computer Society.

[19] Feldbach, M. und K. D. Tönnies: *Robust Line Detection in Historical Church Registers*. In: *Pattern Recognition, 23rd DAGM Symposium*, Seiten 140–147, Munich, Germany, September 2001. Springer-Verlag.

[20] Feldbach, M. und K. D. Tönnies: *Segmentation of the Date in Entries of Historical Church Registers*. In: *Pattern Recognition, 24rd DAGM Symposium*, Seiten 403–410, Zurich, Switzerland, September 2002. Springer-Verlag.

[21] Feldbach, M. und K. D. Tönnies: *Word Segmentation of Handwritten Dates in Historical Documents by Combining Semantic A-Priori-Knowledge with Local Features*. In: *Seventh International Conference on Document Analysis and Recognition*, Seiten 333–337, Edinburgh, UK, August 2003.

[22] Foggia, P., C. Sansone, F. Tortorella und M. Vento: *Combining Statistical and Structural Approaches for Handwritten Character Description*. Image and Vision Computing, 17:701–711, 1999.

[23] Gillies, A. M.: *Cursive Word Recognition Using Hidden Markov Models*. In: *Fifth U.S. Postal Service Advanced Technology Conference*, Seiten 557–562, 1992.

[24] Govindan, V. K. und A. P. Shivaprasad: *Character Recognition – A Review*. Pattern Recognition, 23(7):671–683, 1990.

[25] Graphikon GmbH, Berlin, Germany: *URL:* `www.graphikon.de`.

[26] Grun, P. A.: *Leseschluessel zu unserer alten Schrift*, Band 5 der Reihe *Grundriß der Genealogie*. C. A. Starke Verlag, Limburg, 1984.

[27] Ha, J., R. M. Haralick und I. T. Phillips: *Document Page Decomposition by the Bounding-box Projection Technique*. In: *Third International Conference on Document Analysis and Recognition*, Seiten 1119–1122. IEEE Computer Society, 1995.

[28] Ha, T. und H. Bunke: *Off-line Handwritten Numeral Recognition by Perturbation Method*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(5):535–539, 1997.

[29] Hanmandlu, M., K. R. Murali Mohan und H. Kumar: *Neural Based Handwritten Character Recognition*. In: *Fifth International Conference on Document Analysis and Recognition*, Seiten 241–244, Bangalore, India, 1999.

[30] Heutte, L., T. Paquet, J. V. Moreau, Y. Lecourtier und C. Olivier: *A Structural / Statistical Feature Based Vector for Handwritten Character Recognition*. Pattern Recognition Letters, 19:629–641, 1998.

[31] Hu, Jianming und Hong Yan: *Structural Primitive Extraction and Coding for Handwritten Numeral Recognition*. Pattern Recognition, 31(5):493–509, January 1998.

[32] Jaeger, S., S. Make und A. Waibel: *NPEN++: An On-Line Handwriting Recognition System*. In: *Seventh International Workshop on Frontiers in Handwriting Recognition*, Seiten 249–260, Amsterdam, Netherlands, September 2000.

[33] Kato, Y. und M. Yasuhara: *Recovery of Drawing Order from Scanned Images of Multi-Stroke Handwriting*. In: *Fifth International Conference on Document Analysis and Recognition*, Seiten 705–708, Bangalore, India, September 1999. IEEE Computer Society.

[34] Kavallieratou, E., N. Fakotakis und G. Kokkinakis: *An Unconstrained Handwriting Recognition System*. International Journal on Document Analysis and Recognition, 4:226–242, 2002.

[35] Kazakov, D. und S. Manandhar: *A Hybrid Approach to Word Segmentation*. In: Page, D. (Herausgeber): *Proceedings of the 8th International Conference on Inductive Logic Programming*, Band 1446, Seiten 125–134. Springer-Verlag, 1998.

[36] Keaton, P., H. Greenspan und R. Goodman: *Keyword Spotting for Cursive Document Retrieval*. In: *Workshop on Document Image Analysis*, Seiten 74–81, 1997.

[37] Khan, Nadeem A.: *A Shape Analysis Model with Application to Character and Word Recognition*. Doktorarbeit, Technische Universiteit Eindhoven, Eindhoven, 2000.

[38] Khan, Nadeem A. und Hans A. Hegt: *Recognition of Real-Life Character Samples Using a Structural Variation and Degradation Model*. In: *Fifth International Conference on Document Analysis and Recognition*, Seiten 225–228, Bangalore, India, 1999. IEEE Computer Society.

[39] Kim, G. und V. Govindaraju: *A Lexicon Driven Approach to Handwritten Word Recognition for Real-time Applications*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(4):366–379, April 1997.

[40] Kim, G. und V. Govindaraju: *Handwritten Phrase Recognition as Applied to Street Name Images*. Pattern Recognition, 31(1):41–51, Januar 1998.

[41] Kim, G., V. Govindaraju und S. N. Srihari: *An Architecture for Handwritten Text Recognition Systems*. International Journal on Document Analysis and Recognition, 2(1):37–44, Februar 1999.

[42] Kim, S. H., S. Jeong, G.-S. Lee und C.Y.Suen: *Word Segmentation in Handwritten Korean Text Lines Based on Gap Clustering Techniques*. In: *Sixth International Conference on Document Analysis and Recognition – ICDAR 2001*, Seiten 189–193. IEEE Computer Society, September 2001.

[43] Kimura, F., M. Shridhar und Z. Chen: *Improvements of a Lexicon-directed Algorithm for Recognition of Unconstrained Handwritten Words*. In: *Second International Conference on Document Analysis and Recognition*, Seiten 18–23, Tsukuba Science City, Japan, 1993.

[44] Kolcz, A., J. Alspector, M. Augusteijn, R. Carlson und G. V. Popescu: *A Line-oriented Approach to Word Spotting in Handwritten Documents*. Pattern Analysis and Applications, 3:153–168, 2000.

[45] Kruse, H., R. Mangold, B. Mechler und O. Pengler: *Programmierung Neuronaler Netze: Eine Turbo Pascal Toolbox*. Addison-Wesley, 1991.

[46] Lavrenko, V., T. M. Rath und R. Manmatha: *Holistic Word Recognition for Handwritten Historical Documents*. In: *Proceedings of the International Workshop on Document Image Analysis for Libraries*, Seiten 278–287, Palo Alto, CA, Januar 2004.

[47] Lee, Luan Ling und Natanael Rodrigues Gomes: *Automatic Classification of Deformed Handwritten Numeral Characters*. In: *Fifth International Conference on Document Analysis and Recognition*, Seiten 269–272, Bangalore, India, 1999.

[48] L'Homer, E.: *Extraction of Strokes in Handwritten Characters*. Pattern Recognition, 33(7):1147–1160, 2000.

[49] Lorette, G.: *Handwriting Recognition or Reading? What is the Situation at the Dawn of the 3rd Millenium?* International Journal on Document Analysis and Recognition, 2:2–12, 1999.

[50] Madhvanath, S. und V. Govindaraju: *Local Reference Lines for Handwritten Phrase Recognition*. Pattern Recognition, 32:2021–2028, 1999.

[51] Madhvanath, S. und V. Govindaraju: *The Role of Holistic Paradigms in Handwritten Word Recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(2):149–164, Februar 2001.

[52] Madhvanath, S., E. Kleinberg und V. Govindaraju: *Holistic Verification of Handwritten Phrases*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(12):1344–1356, Dezember 1999.

[53] Madhvanath, S. und V. Kripasundar: *Pruning Large Lexicons Using Generalized Word Shape Descriptors*. In: *Fourth International Conference on Document Analysis and Recognition*, Seiten 552–555, Ulm, Germany, 1997.

[54] Mahadevan, U. und R. C. Nagabushnam: *Gap Metrics for Word Separation in Handwritten Lines.* In: *Third International Conference on Document Analysis and Recognition*, Seiten 124–127, Montreal, Canada, 1995.

[55] Manmatha, R., Chengfeng Han und E. M. Riseman: *Word SPotting: A New Approach to Indexing Handwriting*. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 96*, Seiten 631–637, San Francisco, US, Juni 1996.

[56] Manmatha, R. und N. Srimal: *Scale Space Technique for Word Segmentation in Handwritten Documents*. In: *Scale-Space Theories in Computer Vision*, Seiten 22–33, 1999.

[57] Marti, U. und H. Bunke: *A Full English Sentence Database for Off-Line Handwriting Recognition*. In: *Fifth International Conference on Document Analysis and Recognition*, Seiten 705–708, Bangalore, India, September 1999. IEEE Computer Society.

[58] Marti, U. und H. Bunke: *Text Line Segmentation and Word Recognition in a System for General Writer Independent Handwriting Recognition*. In: *Sixth International Conference on Document*

Analysis and Recognition, Seiten 159–163, Seattle, USA, September 2001. IEEE Computer Society.

[59] Marti, U.-V., R. Messerli und H. Bunke: *Writer Identification Using Text Line Based Features*. In: *Sixth International Conference on Document Analysis and Recognition*, Seiten 101–105, Seattle, USA, September 2001. IEEE Computer Society.

[60] Matic, N., I. Guyon, J. Denker und V. Vapnik: *Writer Adaptation for On-Line Handwritten Character Recognition*. In: *Second International Conference on Document Analysis and Recognition*, Seiten 187–191, Tsukuba Science City, Japan, 1993.

[61] Mazal, Otto: *Lehrbuch der Handschriftenkunde*, Band 10 der Reihe *Elemente des Buch- und Bibliothekswesens*. Dr. Ludwig Reichert Verlag Wiesbaden, 1986.

[62] Michel, Lothar: *Methoden der Forensischen Schriftuntersuchung*. In: Rück, Peter (Herausgeber): *Methoden der Schriftbeschreibung*, Band 4 der Reihe *Historische Hilfswissenschaften*, Seiten 373–386. Jan Thorbecke Verlag, 1999.

[63] Mitchell, T. M.: *Machine Learning*. McGraw-Hill series in computer science. McGraw-Hill, New York, 1997.

[64] Mohamed, M. und P. Gader: *Handwritten Word Recognition Using Segmentation-Free Hidden Markov Modeling and Segmentation-Based Dynamic Programming Techniques*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(5):548–554, 1996.

[65] Morita, M., A. El-Yacoubi, R. Sabourin, F. Bortolozzi und C. Y. Suen: *Handwritten Month Word Recognition on Brazilian Bank Cheques*. In: *Sixth International Conference on Document Analysis and Recognition – ICDAR 2001*, Seiten 972–976. IEEE Computer Society, September 2001.

[66] Morita, M., E. Lethelier, A. El-Yacoubi, F. Bortolozzi und R. Sabourin: *An HMM-Based Approach for Date Recognition*. In: *Fourth IAPR International Workshop on Document Analysis Systems (DAS)*, Seiten 233–244, Rio de Janeiro, Brazil, Dezember 2000.

[67] Morita, Marisa Emika: *Automatic Recognition of Handwritten Dates on Brazilian Bank Cheques*. Doktorarbeit, Université du Québec, Montreal, Canada, Jun 2003.

[68] Nakajima, Y., S. Mori, S. Takegami und S. Sato: *Global Methods for Stroke Segmentation*. International Journal on Document Analysis and Recognition, 2(1):19–23, 1999.

[69] Nielson, H. E. und W. A. Barrett: *Consensus-Based Table Form Recognition*. In: *Seventh International Conference on Document Analysis and Recognition*, Seiten 906–910, Edinburgh, UK, August 2003.

[70] Oh, I.-S. und C. Y. Suen: *Distance Features for Neural Network-Based Recognition of Handwritten Characters*. International Journal on Document Analysis and Recognition, 1(2):73–88, Juli 1998.

[71] Oliveira Jr., J. J. de, J. M. de Carvalho, C. O. de A. Freitas und R. Sabourin: *Feature Sets Evaluation for Handwritten Word Recognition*. In: *Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR 02)*, Seiten 446–451, Niagara-on-the-Lake, Canada, August 2002.

[72] Parisse, C.: *Global Word Shape Processing in Off-Line Recognition of Handwriting*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(4):460–464, April 1996.

[73] Park, J., V. Govindaraju und S. N. Srihari: *Efficient Word Segmentation Driven by Unconstrained Handwritten Phrase Recognition*. In: *Fifth International Conference on Document Analysis and Recognition*, Seiten 605–608, Bangalore, India, 1999. IEEE Computer Society.

[74] Petrucci, A.: *Die Beschriebene Schrift*. In: Rück, Peter (Herausgeber): *Methoden der Schriftbeschreibung*, Band 4 der Reihe *Historische Hilfswissenschaften*, Seiten 9–15. Jan Thorbecke Verlag, 1999.

[75] Plamondon, Réjean und Sargur N. Srihari: *On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(1):63–84, Januar 2000.

[76] Pu, Y. und Z. Shi: *A Natural Learning Algorithm Based on Hough Transform for Text Lines Extraction in Handwritten Documents*. In: *Proceedings of the Sixth International Workshop on Frontiers of Handwriting Recognition (IWFHR VI), Taejon, Korea*, Seiten 637–646, 1998.

[77] Rath, T. M. und R. Manmatha: *Features for Word Spotting in Historical Manuscripts*. In: *Seventh International Conference on Document Analysis and Recognition*, Seiten 218–222, Edinburgh, UK, August 2003.

[78] Rath, T. M., R. Manmatha und V. Lavrenko: *A Search Engine for Historical Manuscript Images*. In: *ACM SIGIR 2004 Conference*, Seiten 369–376, Sheffield, UK, Juli 2004.

[79] Reddy, N. V. Subba und P. Nagabhushan: *A Three-Dimensional Neural Network Model for Unconstrained Handwritten Numeral Recognition: A New Approach*. Pattern Recognition, 31(5):511–516, 1998.

[80] Rocha, J. und T. Pavlidis: *A Shape Analysis Model with Applications to a Character Recognition System*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 16(4):393–404, 1994.

[81] Sayre, K.: *Machine Recognition of Handwritten Words: A Project Report*. Pattern Recognition, 5(3):213–228, 1973.

[82] Schneider, Karin: *Paläographie und Handschriftenkunde für Germanisten: eine Einführung*. Max Niemeyer Verlag, Tübingen, 1999.

[83] Schomaker, L. und E. Segers: *Finding Features Used in the Human Reading of Cursive Handwriting*. International Journal on Document Analysis and Recognition, 2:13–18, 1999.

[84] Seni, G. und E. Cohen: *External Word Segmentation of Off-Line Handwritten Text Lines*. Pattern Recognition, 27(1):41–52, January 1994.

[85] Senior, A. W. und A. J. Robinson: *An Off-Line Cursive Handwriting Recognition System*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(3):309–321, 1998.

[86] Shridhar, M. und F. Kimura: *Segmentation-Based Cursive Handwriting Recognition*. In: Bunke, H. und P. S. P. Wang (Herausgeber): *Handbook of Character Recognition and Document Image Analysis*, Seiten 123–156. World Scientific, Februar 1997.

[87] Steinherz, T., E. Rivlin und N. Intrator: *Offine Cursive Script Word Recognition – a Survey*. International Journal on Document Analysis and Recognition, 2:90–110, 1999.

[88] Tay, Y., P. Lallican, M. Khalid, C. Viard-Gaudin und S. Knerr: *An Offline Cursive Handwritten Word Recognition System*. In: *IEEE Region 10 International Conference on Electrical and Electronic Technology*, Band 2, Seiten 519–524, 2001.

[89] Taylor, I. und M. M. Taylor: *The Psychology of Reading*. New York Academic, 1983.

[90] Teow, Loo-Nin und Kia-Fock Loe: *Robust vision-based feature and classification schemes for off-line handwritten digit recognition*. Pattern Recognition, 35:2355–2364, 2002.

[91] Tomai, C. I., B. Zhang und V. Govindaraju: *Transcript Mapping for Historic Handwritten Document Images*. In: *Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR 02)*, Seiten 413–418, Niagara-on-the-Lake, Canada, August 2002.

[92] Verein für Computergenealogie e.V.: *c/o K.-P. Wessel, Lampehof 58, 28259 Bremen. URL:* `www.genealogienetz.de/vereine/CompGen/`.

[93] Vinciarelli, A.: *A Survey on Off-line Cursive Word Recognition*. Pattern Recognition, 35:1433–1446, 2002.

[94] Vinciarelli, A. und S. Bengio: *Writer Adaptation Techniques in Off-Line Cursive Word Recognition*. In: *Eighth International Workshop on Frontiers in Handwriting Recognition*, Seiten 287–291, Ontario, Canada, 2002.

[95] Vinciarelli, A. und J. Luettin: *A new normalization technique for cursive handwritten words*. Pattern Recognition Letters, 22(9):1043–1050, 2001.

[96] Vuori, V., J. Laaksonen, E. Oja und J. Kangas: *Experiments with Adaptation Strategies for a Prototype-Based Recognition System for Isolated Handwritten Characters*. International Journal on Document Analysis and Recognition, 3:150–159, 2001.

[97] Vuurpijl, L. und L. Schomaker: *Coarse Writing-style Clustering Based on Simple Stroke-related Features*. In: Downton, A. C. und S. Impedovo (Herausgeber): *Progress in Handwriting Recognition*, Seiten 37–44. World Scientific Publishers, 1997.

[98] Wang, L. und T. Pavlidis: *Direct Gray-Scale Extraction of Features for Character Recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 15(10):1053–1067, Oktober 1993.

[99] Warwel, Kurt: *Die Vereinfachte Ausgangsschrift (VA) als Konsequenz der Schulschriftentwicklung*. In: Rück, Peter (Herausgeber): *Methoden der Schriftbeschreibung*, Band 4 der Reihe *Historische Hilfswissenschaften*, Seiten 469–479. Jan Thorbecke Verlag, 1999.

[100] Wienecke, M., G. A. Fink und G. Sagerer: *A Handwriting Recognition System Based on Visual Input*. In: *Second International Workshop on Computer Vision Systems*, Seiten 63–72, Vacouver, Canada, 2001.

[101] Xiao, X. und G. Leedham: *Cursive Script Segmentation Incorporating Knowledge of Writing*. In: *Fifth International Conference on Document Analysis and Recognition*, Seiten 535–538, Bangalore, India, 1999.

[102] Xu, Q., L. Lam und C. Y. Suen: *Recognition of Handwritten Month Words on Cheques*. In: *Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR 02)*, Seiten 111–116, Ontario, Canada, August 2002.

[103] Zimmer, Alf C.: *Argumente für die Bedeutung der Impliziten Dynamik beim Lesen Handgeschriebener Texte – Experimentelle Daten und ein Theoretisches Modell*. In: Rück, Peter (Herausgeber): *Methoden der Schriftbeschreibung*, Band 4 der Reihe *Historische Hilfswissenschaften*, Seiten 453–461. Jan Thorbecke Verlag, 1999.

[104] Zimmermann, M. und J. Mao: *Lexicon Reduction Using Key Characters in Cursive Handwritten Words*. Pattern Recognition Letters, 20:1297–1304, 1999.

Anhang A

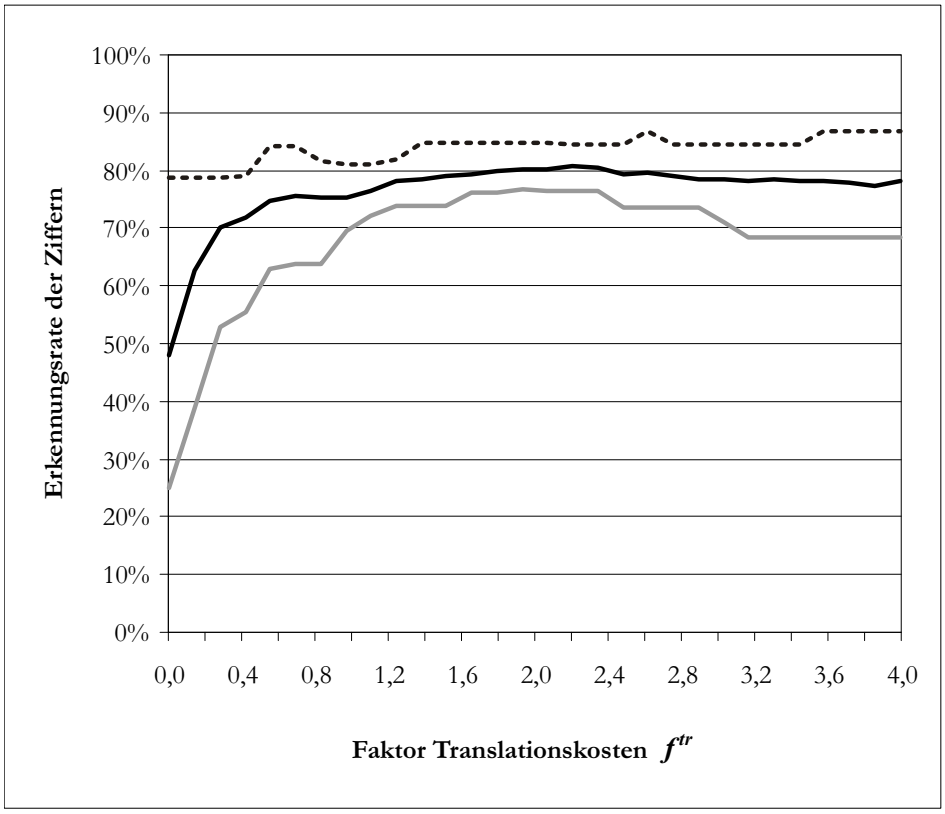
Parameterabhängigkeit

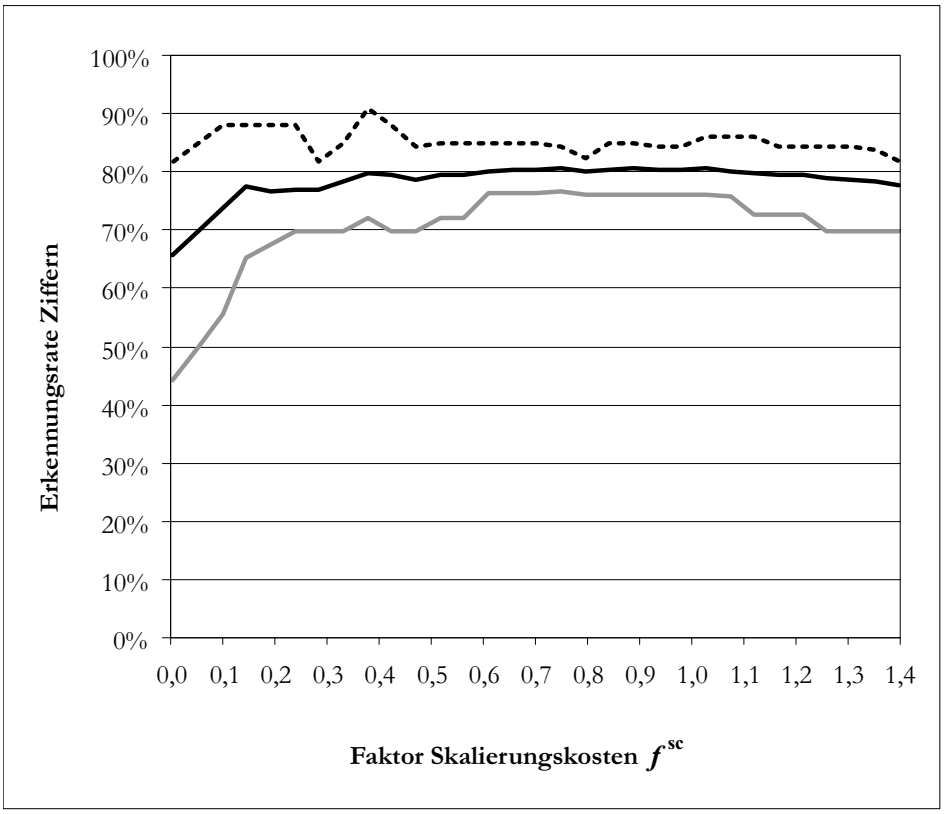
In diesem Anhang sind die Diagramme über den Verlauf der Erkennungsrate in Abhängigkeit der Kostenfaktoren der einzelnen Abweichungsarten sowie weiterer Parameter dargestellt. Die drei Kurven repräsentieren die Erkennungsrate der besten Klasse (gestrichelt), der schlechtesten Klasse (grau) und die mittlere Erkennungsrate (schwarz).

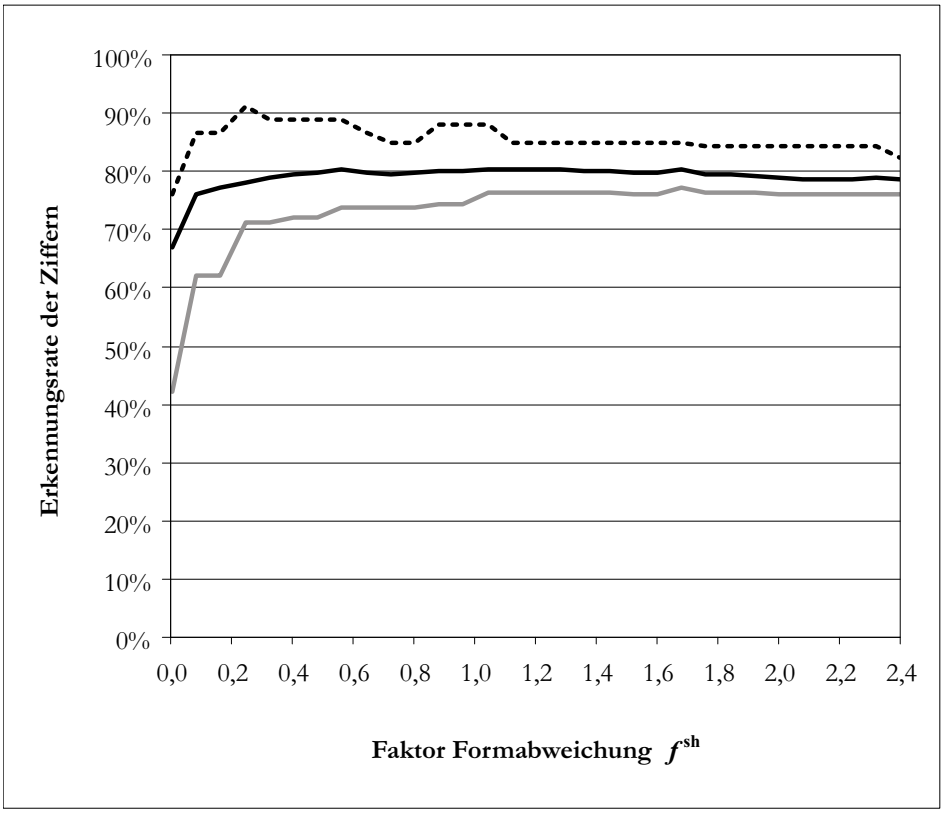
Der erste Abschnitt bezieht sich auf die Faktoren des Matchings. Die Tests wurden auf der Basis der Ziffernerkennung durchgeführt.

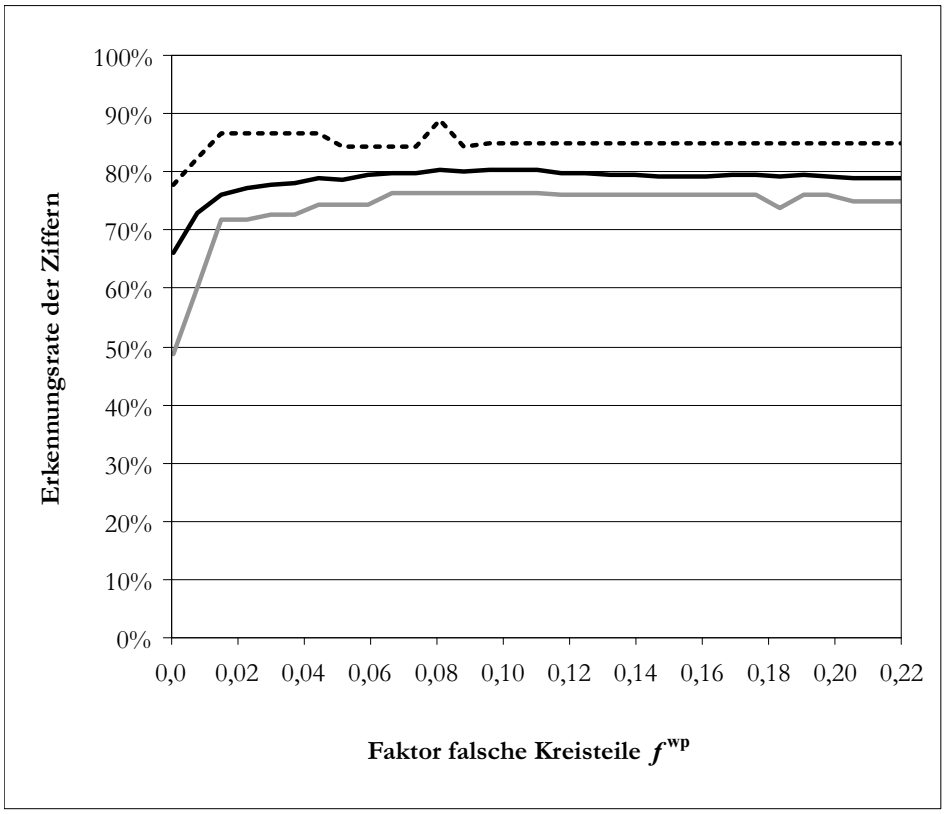
Im zweiten Abschnitt sind die Ergebnisse der Parameter-Tests für die Worterkennung aufgeführt.

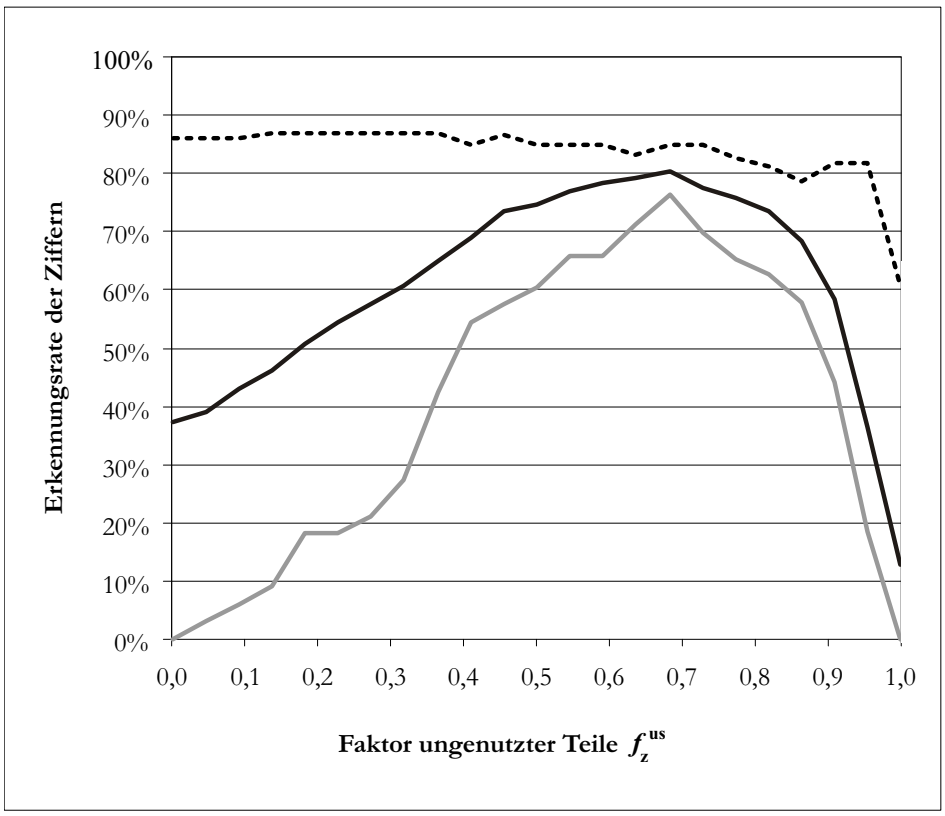
A.1 Faktoren des Matchings zwischen Primitiven

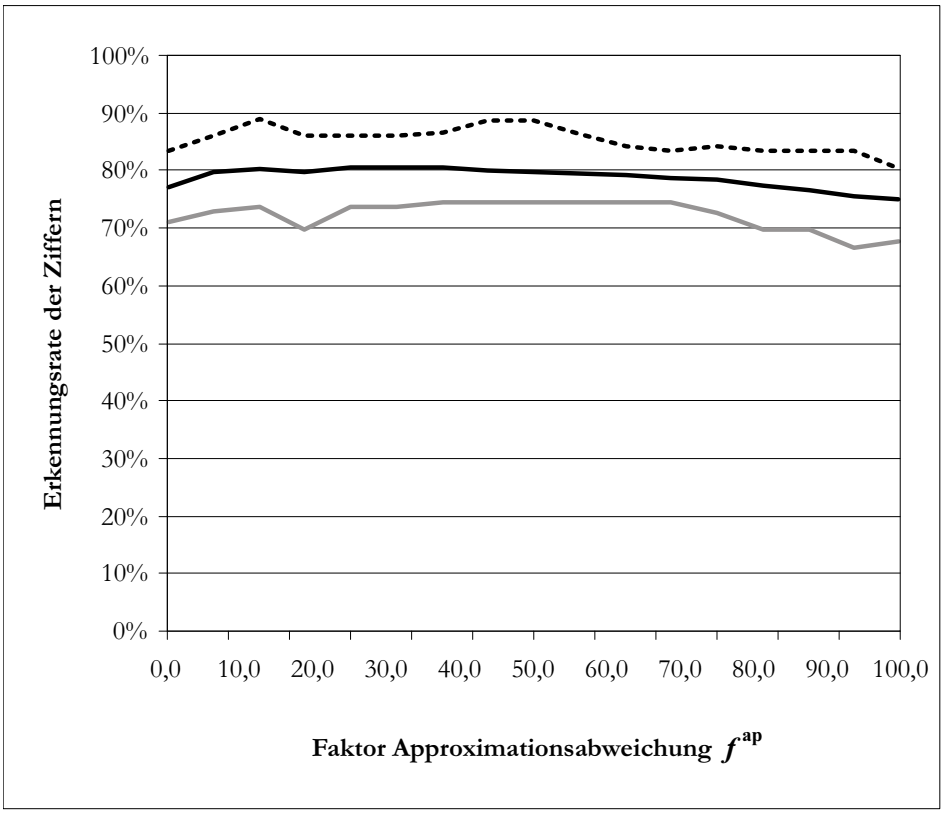



100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Erkennungsrate Ziffern
0,0 0,1 0,2 0,3 0,4 0,5 0,6 0,7 0,8 0,9 1,0 1,1 1,2 1,3 1,4
Faktor Skalierungskosten f^{sc}


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Erkennungsrate der Ziffern
0,0 0,2 0,4 0,6 0,8 1,0 1,2 1,4 1,6 1,8 2,0 2,2 2,4
Faktor Formabweichung f^{sh}

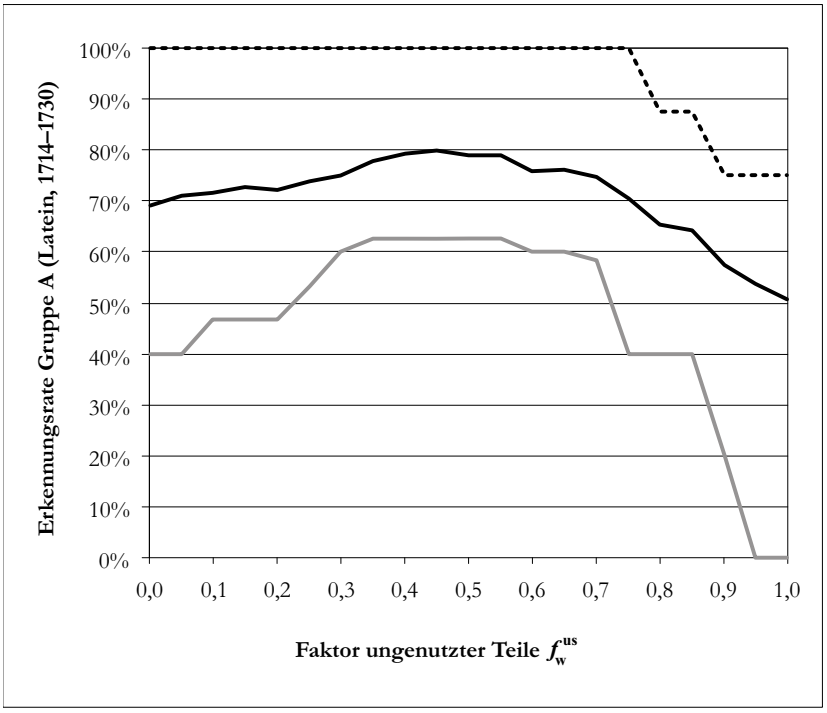

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Erkennungsrate der Ziffern
0,0 0,02 0,04 0,06 0,08 0,10 0,12 0,14 0,16 0,18 0,20 0,22
Faktor falsche Kreisteile f^{wp}

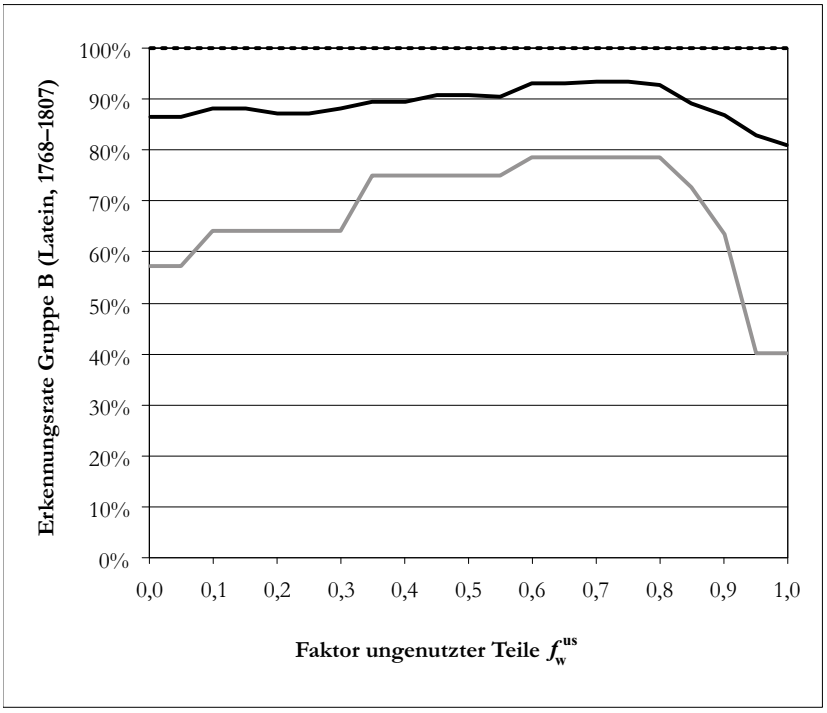

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Erkennungsrate der Ziffern
0,0 0,1 0,2 0,3 0,4 0,5 0,6 0,7 0,8 0,9 1,0
Faktor ungenutzter Teile f_z^{us}

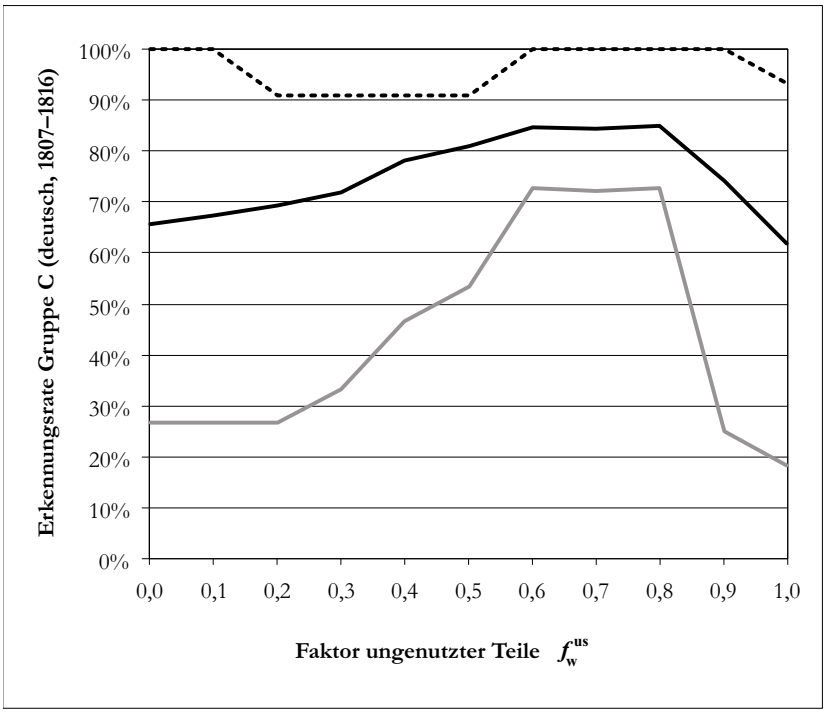

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Erkennungsrate der Ziffern
0,0
10,0
20,0
30,0
40,0
50,0
60,0
70,0
80,0
90,0
100,0
Faktor Approximationsabweichung f^{ap}

A.2 Parameter der Worterkennung

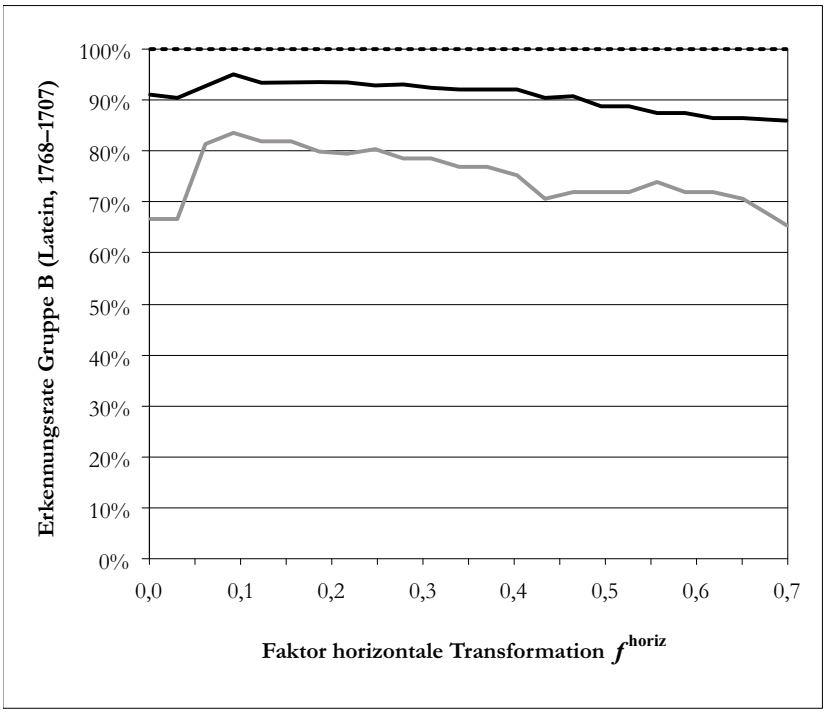
Einfluss des Faktors $f_{\mathrm{w}}^{\mathrm{us}}$ auf die drei Schriftstilgruppen:

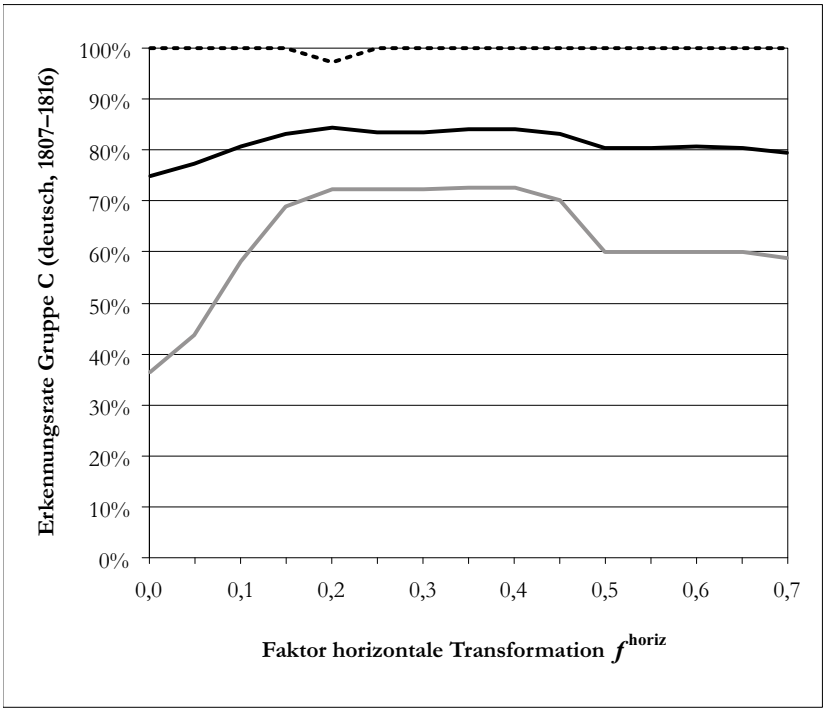




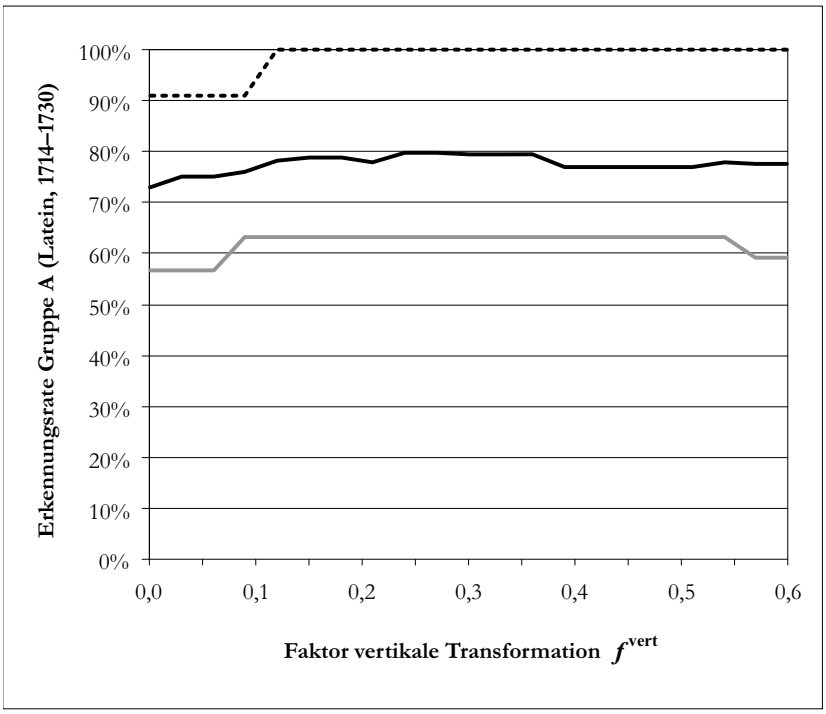


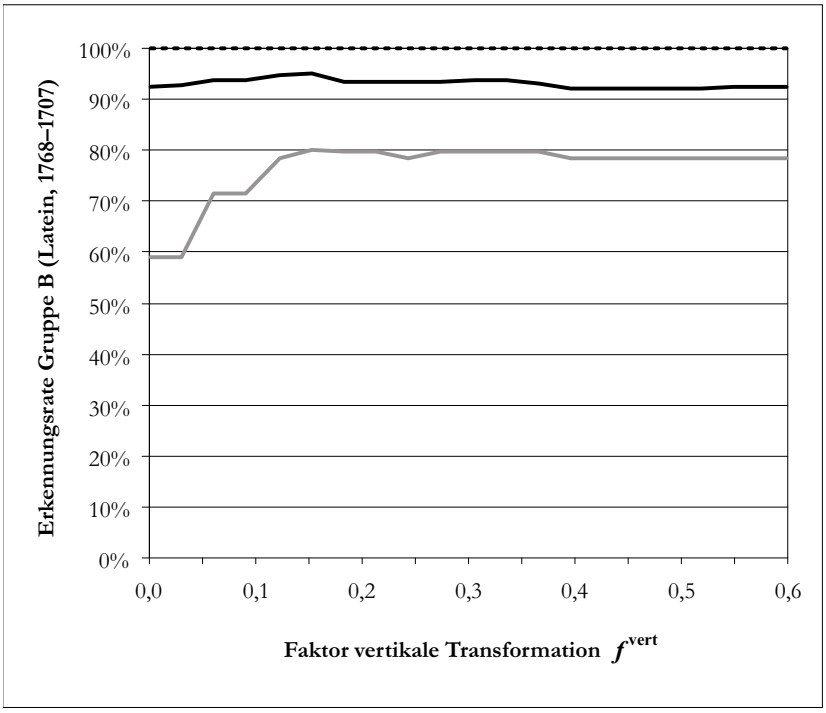
Einfluss des Faktors f^{horiz} auf die drei Schriftstilgruppen:

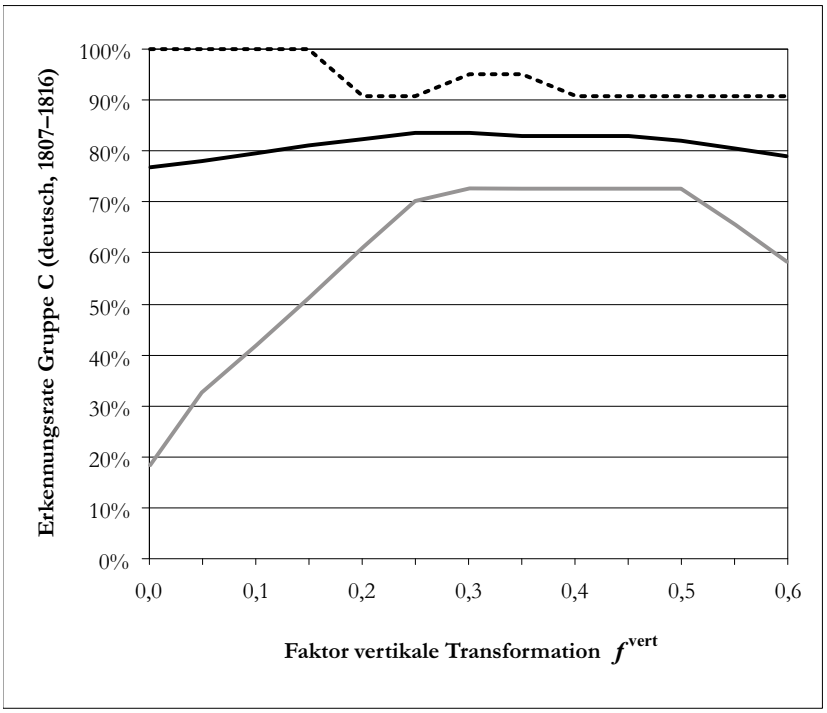

Erkennungsrate Gruppe B (Latein, 1768–1707)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0,0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
Faktor horizontale Transformation f^{horiz}


Erkennungsrate Gruppe C (deutsch, 1807–1816)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0,0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
Faktor horizontale Transformation f^{horiz}

Einfluss des Faktors f^{vert} auf die drei Schriftstilgruppen:






100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Erkennungsrate Gruppe C (deutsch, 1807–1816)
0,0
0,1
0,2
0,3
0,4
0,5
0,6
Faktor vertikale Transformation f^{vert}

Anhang B

Prototypen

Ziffer	Prototypen
0	
1	
2	
3	
4	
5	
6	
7	
8	
9	

Abbildung B.1: Prototypen der Ziffernerkennung.

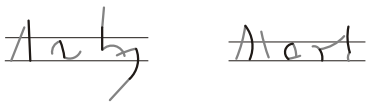
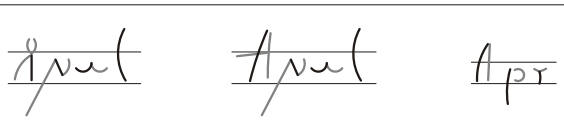
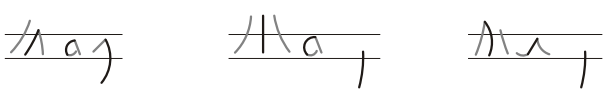
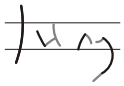
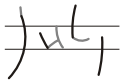
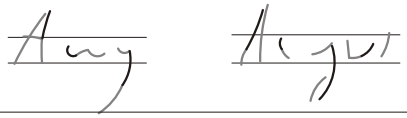
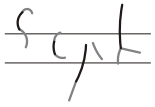
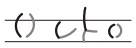
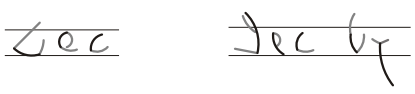
Monat	Prototypen
Januar	
Februar	
März	
April	
Mai	
Juni	
Juli	
August	
September	
Oktober	
November	
Dezember	

Abbildung B.2: Prototypen der lateinischen Monatsnamen aus Wegenstedt, 1714–1730.

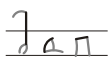
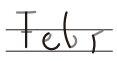
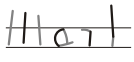
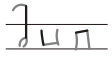
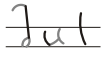
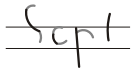
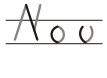
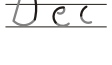
Monat	Prototypen
Januar	
Februar	
März	
April	
Mai	
Juni	
Juli	
August	
September	
Oktober	
November	
Dezember	

Abbildung B.3: Prototypen der lateinischen Monatsnamen aus Wegenstedt, 1768–1807.

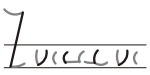
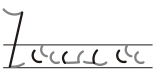
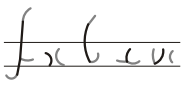
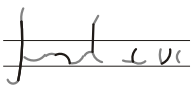
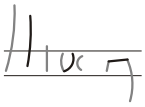
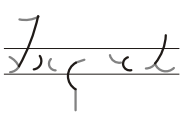
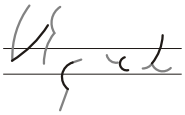
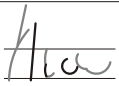
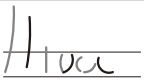
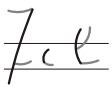
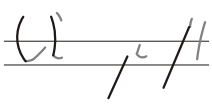
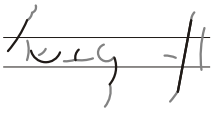
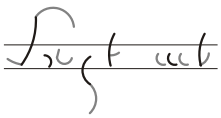
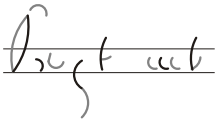
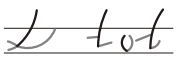
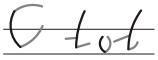
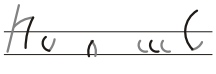
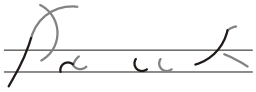
Monat	Prototypen	
Januar		
Februar		
März		
April		
Mai		
Juni		
Juli		
August		
September		
Oktober		
November		
Dezember		

Abbildung B.4: Prototypen der deutschen Monatsnamen aus Wegenstedt, 1807–1816.