

Stefan Fleischer

Design und Implementierung eines
Multi-Classifer-Systems (MCS) für die
Erkennung von gerendertem Text

Diplomarbeit

BEI GRIN MACHT SICH IHR WISSEN BEZAHLT



- Wir veröffentlichen Ihre Hausarbeit, Bachelor- und Masterarbeit
- Ihr eigenes eBook und Buch - weltweit in allen wichtigen Shops
- Verdienen Sie an jedem Verkauf

Jetzt bei www.GRIN.com hochladen
und kostenlos publizieren



Bibliografische Information der Deutschen Nationalbibliothek:

Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de/> abrufbar.

Dieses Werk sowie alle darin enthaltenen einzelnen Beiträge und Abbildungen sind urheberrechtlich geschützt. Jede Verwertung, die nicht ausdrücklich vom Urheberrechtsschutz zugelassen ist, bedarf der vorherigen Zustimmung des Verlanges. Das gilt insbesondere für Vervielfältigungen, Bearbeitungen, Übersetzungen, Mikroverfilmungen, Auswertungen durch Datenbanken und für die Einspeicherung und Verarbeitung in elektronische Systeme. Alle Rechte, auch die des auszugsweisen Nachdrucks, der fotomechanischen Wiedergabe (einschließlich Mikrokopie) sowie der Auswertung durch Datenbanken oder ähnliche Einrichtungen, vorbehalten.

Impressum:

Copyright © 2007 GRIN Verlag
ISBN: 9783640200832

Dieses Buch bei GRIN:

<https://www.grin.com/document/117720>

Stefan Fleischer

Design und Implementierung eines Multi-Classifizier-Systems (MCS) für die Erkennung von gerendertem Text

GRIN - Your knowledge has value

Der GRIN Verlag publiziert seit 1998 wissenschaftliche Arbeiten von Studenten, Hochschullehrern und anderen Akademikern als eBook und gedrucktes Buch. Die Verlagswebsite www.grin.com ist die ideale Plattform zur Veröffentlichung von Hausarbeiten, Abschlussarbeiten, wissenschaftlichen Aufsätzen, Dissertationen und Fachbüchern.

Besuchen Sie uns im Internet:

<http://www.grin.com/>

<http://www.facebook.com/grincom>

http://www.twitter.com/grin_com

Diplomarbeit

**Design und Implementierung eines
Multi-Classifer-Systems (MCS)
für die Erkennung von
gerendertem Text**

vorgelegt von

Stefan Fleischer

August 2007

Westfälische Wilhelms-Universität Münster

Institut für Informatik

Computer Vision and Pattern Recognition Group

Inhaltsverzeichnis

1	Einleitung	1
2	Erkennung von gerendertem Text	3
2.1	Eigenschaften gerenderten Textes	3
2.2	Bisheriges OCR-System	6
2.2.1	Vorverarbeitung	6
2.2.2	Hybride Klassifikation	8
2.2.3	Nachbearbeitung	13
2.3	Aktueller Stand und Optimierungsmöglichkeiten	13
3	Lern- und Testdaten	20
3.1	Kategorisierung gerendeter Texte	20
3.2	Format der Datenbanken	23
3.3	Erweiterung der Datenbasis	24
4	Konstruktion und Test einzelner Klassifikatoren	29
4.1	Konstruktion von Klassifikatoren	29
4.2	Analyse und Vergleich von Klassifikatoren	33
4.3	Erzielte Testergebnisse	40
4.3.1	Klassifikatortests	40
4.3.2	Systemtests	46
5	Konstruktion und Test von Multi-Classifer-Systemen	56
5.1	Ansätze zur Kombination mehrerer Klassifikatoren	56
5.2	Erzielte Testergebnisse	64
5.2.1	Klassifikatortests	65
5.2.2	Systemtests	72
6	Fazit und Ausblick	80
A	Trainingsdaten der Version 2006	82
A.1	Lern- und Testdaten der Screen-Char-Datenbank	82
A.2	Testdaten der Screen-Word-Datenbank	85
A.3	Format der alten Datenbanken	86
B	Trainingsdaten der Version 2007-MCS	87
B.1	Lern- und Testdaten der Screen-Char-Datenbank	87
B.2	Testdaten der Screen-Word-Datenbank	88
C	Resultate durchgeführter Testläufe	89
C.1	Resultate durchgeführter Klassifikatortests	89
C.2	Resultate durchgeführter Systemtests	92

Abbildungsverzeichnis

2.1	Beispiele verschiedener Rendering-Techniken	4
2.2	Schwierig zu segmentierende Wörter	5
2.3	Auswirkungen geringer Änderungen des Zooms	5
2.4	OCR-Prozess des bestehenden Systems zur Erkennung von gerendertem Text	7
2.5	Schwellwertkorrektur bei der Textsegmentierung	8
2.6	Segmentierung eines gerenderten Wortes	9
2.7	Wirkung des Kostenterms c_{nb}	10
2.8	Hypothesengraph für das Beispielwort 'arm'	13
2.9	Schwierige Wahl geeigneter Parameter für die Wortsegmentierung	15
2.10	Beispiele schwierig zu segmentierender Buchstabenkombinationen	15
3.1	Veranschaulichung der Kategorien neuer Screen-Chars	25
4.1	Normierung der Aspektwerte	33
4.2	Unterschiedliche Klassifikationen des Buchstabens 'F'	43
4.3	Abhängigkeit der Erkennungsrate von der Zoning-Rastergröße bei Merkmalen auf Basis der Ableitung	46
4.4	Beispiele vom Klassifikator <code>knn_5x5m</code> nicht korrekt erkannter Screen-Words geringer x-Höhe	54
4.5	Serifen unterstützen die Erkennung falsch segmentierter Buchstaben	55
5.1	Kombination von Typ-3-Klassifikatoren über Aggregationsfunktionen	60

Tabellenverzeichnis

3.1	Häufigkeiten einzelner Kategorien der Screen-Chars	26
3.2	Häufigkeiten einzelner Kategorien der Screen-Words	27
4.1	Konstruierte Klassifikatoren und deren Merkmalsgruppen	34
4.2	Beispiel einer Confusion-Matrix	37
4.3	Beispielhaftes Ergebnis zweier Klassifikatoren	38
4.4	Ausschnitt der Confusion-Matrix des Klassifikators $\mathbf{knn_5x5m}$	41
4.5	Resultate: Klassifikatortests einzelner Klassifikatoren — $\mathbf{knn_zxx(e m)}$, mit $z = 5, \dots, 10$	42
4.6	Resultate: Klassifikatortests einzelner Klassifikatoren — $\mathbf{knn_zxx(e m)a}$, mit $z = 5, \dots, 10$	44
4.7	Resultate: Klassifikatortests einzelner Klassifikatoren — $\mathbf{knn_1 2 3s_zxx(e m)}$, mit $z = 5, \dots, 10$	45
4.8	Falsch separierte Screen-Words nach x-Höhe und Serifen kategorisiert	47
4.9	Erzielte Erkennungsraten von Wörtern verschiedener x-Höhen des Klassifikators $\mathbf{knn_5x5m}$	48
4.10	Resultate: Systemtests einzelner Klassifikatoren — $\mathbf{knn_zxx(e m)}$, mit $z = 5, \dots, 10$	49
4.11	Resultate: Systemtests einzelner Klassifikatoren — $\mathbf{knn_zxx(e m)a}$, mit $z = 5, \dots, 10$	50
4.12	x^2 -Werte von McNemar-Tests bzgl. der Verbesserung durch das Höhen-Breitenverhältnis als zusätzliches Merkmal	51
4.13	Resultate: Systemtests einzelner Klassifikatoren — $\mathbf{knn_1s_zxx(e m)}$, mit $z = 5, \dots, 10$	52
4.14	Resultate: Systemtests einzelner Klassifikatoren — $\mathbf{knn_2s_zxx(e m)}$, mit $z = 5, \dots, 10$	53
4.15	Resultate: Systemtests einzelner Klassifikatoren — $\mathbf{knn_3s_zxx(e m)}$, mit $z = 5, \dots, 10$	53
4.16	Vom Klassifikator $\mathbf{knn_5x5m}$ korrekt erkannte serifenbehaftete und serifenlose Screen-Words	55
5.1	Beispiel der Möglichkeit zur Überschreitung der theoretischen oberen Grenze	64
5.2	Resultate: Klassifikatortests einzelner Klassifikatoren — $\mathbf{knn_zxx(e m)}$, mit $z = 5, \dots, 10$	65
5.3	Übereinstimmende Fehlklassifikationen der zwei Gruppen $\mathbf{knn_zxx(e m)}$, mit $z = 5, \dots, 10$	66
5.4	Auswirkungen der Kombination auf den Fehler der zwei Gruppen $\mathbf{knn_zxx(e m)}$, mit $z = 5, \dots, 10$	67
5.5	Resultate: Klassifikatortests kombinierter Klassifikatoren — $\mathbf{knn_zxx(e m)a}$, mit $z = 5, \dots, 10$	67
5.6	Auswirkungen der Kombination auf den Fehler der zwei Gruppen $\mathbf{knn_zxx(e m)a}$, mit $z = 5, \dots, 10$	68
5.7	Übereinstimmende Fehlklassifikationen der Gruppe $\mathbf{knn_2s_zxxm}$	69
5.8	CFD-Werte basierend auf Klassifikatortests	69
5.9	Resultate: Klassifikatortests kombinierter Klassifikatoren — $\mathbf{knn_1s_zxx(e m)}$, mit $z = 5, \dots, 10$	70

5.10	Resultate: Klassifikatortests kombinierter Klassifikatoren — $\text{knn_2szxz}(\mathbf{e m})$, mit $z = 5, \dots, 10$	70
5.11	Resultate: Klassifikatortests kombinierter Klassifikatoren — $\text{knn_3szxz}(\mathbf{e m})$, mit $z = 5, \dots, 10$	70
5.12	Auswirkungen der Kombination auf den Fehler der zwei Gruppen $\text{knn_1szxz}(\mathbf{e m})$, mit $z = 5, \dots, 10$	71
5.13	Auswirkungen der Kombination auf den Fehler der zwei Gruppen $\text{knn_2szxz}(\mathbf{e m})$, mit $z = 5, \dots, 10$	71
5.14	Auswirkungen der Kombination auf den Fehler der zwei Gruppen $\text{knn_3szxz}(\mathbf{e m})$, mit $z = 5, \dots, 10$	71
5.15	Resultate: Systemtests kombinierter Klassifikatoren — $\text{knn_zxx}(\mathbf{e m})$, mit $z = 5, \dots, 10$	74
5.16	Resultate: Systemtests kombinierter Klassifikatoren — $\text{knn_zxx}(\mathbf{e m})\mathbf{a}$, mit $z = 5, \dots, 10$	75
5.17	Resultate: Systemtests kombinierter Klassifikatoren $\text{knn_1szxz}(\mathbf{e m})$, mit $z = 5, \dots, 10$	76
5.18	Resultate: Klassifikatortests kombinierter Klassifikatoren — $\text{knn_2szxz}(\mathbf{e m})$, mit $z = 5, \dots, 10$	77
5.19	Resultate: Systemtests kombinierter Klassifikatoren — $\text{knn_3szxz}(\mathbf{e m})$, mit $z = 5, \dots, 10$	78
A.1	Häufigkeiten einzelner Kategorien von Screen-Chars der Version 2006 (Schriftstil „normal“)	82
A.2	Häufigkeiten einzelner Kategorien von Screen-Chars der Version 2006 (Schriftstil „fett“)	83
A.3	Häufigkeiten einzelner Kategorien von Screen-Chars der Version 2006 (Schriftstil „kursiv“)	83
A.4	Häufigkeiten einzelner Kategorien von Screen-Chars der Version 2006 (Schriftstil „fett-kursiv“)	84
A.5	Häufigkeiten einzelner Kategorien von Screen-Words der Version 2006	85
A.6	Format der alten Screen-Char-Datenbank	86
A.7	Format der alten Screen-Word-Datenbank	86
B.1	Häufigkeiten einzelner Kategorien von Screen-Chars der Version 2007-MCS	87
B.2	Häufigkeiten einzelner Kategorien von Screen-Words der Version 2007-MCS	88
C.1	Resultate von Klassifikatortests der zwei Gruppen $\text{knn_zxx}(\mathbf{e m})$, mit $z = 5, \dots, 10$.	89
C.2	Resultate von Klassifikatortests der zwei Gruppen $\text{knn_zxx}(\mathbf{e m})\mathbf{a}$, mit $z = 5, \dots, 10$	90
C.3	Resultate von Klassifikatortests der zwei Gruppen $\text{knn_1szxz}(\mathbf{e m})$, mit $z = 5, \dots, 10$	90
C.4	Resultate von Klassifikatortests der zwei Gruppen $\text{knn_2szxz}(\mathbf{e m})$, mit $z = 5, \dots, 10$	91
C.5	Resultate von Klassifikatortests der zwei Gruppen $\text{knn_3szxz}(\mathbf{e m})$, mit $z = 5, \dots, 10$	91
C.6	Resultate von Systemtests der Gruppe knn_zxxe , mit $z = 5, \dots, 10$	92
C.7	Resultate von Systemtests der Gruppe knn_zxxm , mit $z = 5, \dots, 10$	93
C.8	Resultate von Systemtests der Gruppe knn_zxxea , mit $z = 5, \dots, 10$	93
C.9	Resultate von Systemtests der Gruppe knn_zxxma , mit $z = 5, \dots, 10$	94
C.10	Resultate von Systemtests der Gruppe knn_1szxxe , mit $z = 5, \dots, 10$	94
C.11	Resultate von Systemtests der Gruppe knn_1szxxm , mit $z = 5, \dots, 10$	95
C.12	Resultate von Systemtests der Gruppe knn_2szxxe , mit $z = 5, \dots, 10$	95
C.13	Resultate von Systemtests der Gruppe knn_2szxxm , mit $z = 5, \dots, 10$	96
C.14	Resultate von Systemtests der Gruppe knn_3szxxe , mit $z = 5, \dots, 10$	96
C.15	Resultate von Systemtests der Gruppe knn_3szxxm , mit $z = 5, \dots, 10$	97

Kapitel 1

Einleitung

Die Einführung elektronischer Textverarbeitung führte in vielerlei Hinsicht zu immensen Erleichterungen und Effizienzsteigerungen. Texte können schnell und einfach geändert, kopiert, gelöscht oder zur weiteren Verarbeitung in andere Umgebungen überführt werden. Die Textverarbeitung wird dabei von Anwendungen zur Visualisierung der Schrift und zur Entgegennahme von Eingaben unterstützt, wobei die Textinhalte gelöst von ihrer Darstellung und verschiedenen Anwendungsfällen als Zeichenketten kodiert vorliegen.

Doch oft ist auf pixelbasierten Ausgabegeräten nur die grafische Repräsentation von Zeichenketten vorhanden. Die Beschriftungen von Anwendungsfenstern und sonstigen Steuerelementen können i.d.R. noch beim Betriebssystem erfragt werden. Die Inhalte geschützter Dokumente und Texte in Pixelgrafiken sind demgegenüber allerdings nur optisch vorhanden. Letztere liegen nicht mal mehr versteckt als Zeichenketten kodiert vor. Die dargestellten Zeichen lassen sich also nur mittels OCR (Optical Character Recognition, optische Zeichenerkennung oder auch automatische Texterkennung) ermitteln.

An die Zeichenerkennung bei pixelbasierten Ausgabegeräten mit ihrer relativ geringen Auflösung und groben Rasterung werden andere Herausforderungen gestellt als bei der Verarbeitung eingescannter Texte. Verbreitete Ansätze zur klassischen Erkennung eingescannter Texte lassen sich daher nur teilweise auf gerenderte Texte übertragen und dort sinnvoll nutzen. Deshalb bedarf es ausgefeilter Techniken, die den gestellten Herausforderungen gewachsen sind.

Am Institut für Informatik der Westfälischen Wilhelms-Universität Münster wurde von Steffen Wachenfeld im Rahmen einer Doktorarbeit und Hans-Ulrich Klein im Rahmen einer Diplomarbeit ein OCR-System entwickelt, das auf die Erkennung gerendertexte spezialisiert ist. Die ersten Tests dieses OCR-Systems sind vielversprechend, das Konzept erweist sich als innovativer Ansatz mit hohem Potenzial.

Es ist typisch für Innovationsprozesse, dass sie eine Reihe von inkrementellen Entwicklungsphasen durchlaufen. *Ziel dieser Arbeit* ist es, die Klassifikationskomponente des am Institut für Informatik entwickelten OCR-Systems zur Erkennung von gerendertem Text durch die Implementierung eines Multi-Classifer-Systems (MCS) weiter zu optimieren.

Die Klassifikationskomponente klassifiziert die gerenderten Schriften und weist ihnen somit die erkannten Textzeichen zu. Die Wahl eines besten Klassifikators für diese Aufgabe stellt sich als