Stefan Fleischer

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

Diplomarbeit

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.

Impressum:

Copyright © 2007 GRIN Verlag
ISBN: 9783640200832

Stefan Fleischer

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

GRIN Verlag

Diplomarbeit

Design und Implementierung eines Multi-Classifier-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

Abbildungsverzeichnis

Tabellenverzeichnis

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 gerenderter Texte 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-Classifier-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