

Document Image Analysis

Prof. Dr. Christoph Dalitz
Sommersemester 2011

Vorlesung an der
Hochschule Niederrhein

1

1 Überblick (2)

Zwei Dokumentkategorien bei der Analyse:

- Text
typische Aufgaben: Ermitteln Spalten und Zeilen,
Erkennung der einzelnen Buchstaben (OCR)
- Grafik
typische Aufgaben: Vektorisierung, Erkennen der
grafischen Primitive (Linien, Rechtecke, ...)

Nicht Gegenstand der Document Image Analysis:

- Bildanalyse, z.B. von medizinischen Bildern,
Erkennung bewegter Objekte oder Gesichtserkennung
- dazu Veranstaltung im Master Informatik (Frau Pohle)

3

1 Überblick (1)

Zielsetzung der Document Image Analysis:

- Gewinnung der semantischen Information
· aus dem Bild eines Dokuments

□ Beispiele

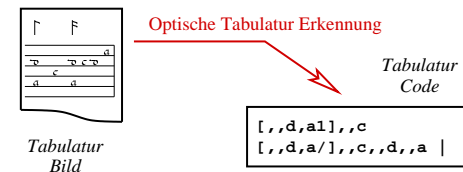
- Gewinnung des Textes in Lesereihenfolge
- Erkennung eines elektrischen Schaltbildes
- Optische Musikererkennung

2

1 Überblick (3)

Beispielsystem (an HN entwickelt):

Optische Erkennung historischer Tabulaturdrucke
(siehe <http://otr4gamera.sourceforge.net/>)



Ziele dieser Veranstaltung:

- verstehen wie solch ein Systems funktioniert
- eigene DIA-Systeme mithilfe eines entsprechenden Frameworks erstellen können

4

1 Überblick (4)

Praktische Übungen

- Themen der Vorlesung werden unmittelbar danach in der Übung am Rechner ausprobiert
- dabei wird das *Gamera-Framework* eingesetzt
 - Python-Bibliothek ⇒ Python Crashkurs in Veranstaltung
 - DIA-Funktionen direkt am Beispiel erläutert

Voraussetzung:

- Grundkenntnisse im Programmieren (EPR)

5

Literatur

Bücher zur DIA lediglich lose Artikelsammlungen.
Frei zugänglicher Überblick als Einstieg:

- R. Kasturi, L. O'Gorman, V. Govindaraju: *Document image analysis: A primer*. Sadhana Vol. 27, Part 1, February 2002, pp. 3-22. Siehe <http://www.ias.ac.in/sadhana/>

Gamera Homepage:

- <http://gamera.sf.net/>

Gute Einführung und Referenz zu Python:

- D.M. Beazley: *Python - Essential Reference*. New Riders 2001

7

1 Überblick (5)

Gamera Framework

- an Johns Hopkins University (USA) entwickelt und unter der GNU General Public License bereitgestellt; Komponenten auch an unserer Hochschule entstanden
- Python-Bibliothek zur Erstellung von Dokument-Analysissystemen durch "Domain-Experts"
- Plattformunabhängig mit (optionaler) wxPython-Oberfläche

Neben eingebauten Routinen bietet Gamera

- *Plugin-Mechanismus* um zusätzliche Python-Routinen in C++ zu implementieren
- *Toolkit-Konzept* um Plugins und Lösungen zu optional ladbaren Bibliotheken zusammenzufassen

6

Themen (1)

2 Grundbegriffe

- Bildtypen und -formate, Vektorformat versus Bitmap
- Bildrepräsentation in Gamera
- Binarisation, Kombination von Onebit-Bildern
- Projektionen, Runlength, Connected Components
- typischer Aufbau eines DIA-Systems

3 Preprocessing

- Noise Reduction (Filter, Morphologie, kFill)
- Skew Correction
- Charakteristische Dimensionen

8

Themen (2)

4 Segmentation

- Page-Segmentation
- Evaluation der Segmentierungsqualität
- Connected-Component Analyse

5 Optical Character Recognition (OCR)

- Allgemeines Schema
- Begriff des "Features" und das kNN-Verfahren
- Beispiele für Features
- Methoden zur Nachkorrektur

9

Themen (3)

6 Grafische Dokumente

- Text-Graphics Separation
- Outline, Thinning und Skeletonization
- Linien- und Kreisbögen-Erkennung
- Linien und Kurven-Fitting
- Polygonisierung und kritische Punkte

10