CSL Cognitive Systems Laboratory

Manuskriptanalyse mit dem Computer

Bernd Neumann

Arbeitsbereich Kognitive Systeme
Department Informatik
Fakultät für Mathematik, Informatik und Naturwissenschaften
Universität Hamburg

Dank

Prof. Michael Friedrich
Asien-Afrika-Institut
Abt. für Sprache und Kultur Chinas

Dr. Matthias Richter
The University of Chicago
Department of East Asian Languages and Civilizations

Agenda

- Schreibstil in der Manuskriptforschung
- Kleine Einführung in die digitale Bildverarbeitung
- Layout-Analyse
- Strukturelle Zeichenanalyse
- Clustering und Datamining

Schreibstil in der Manuskriptforschung

Richter 2006:
"Tentative Criteria for Discerning Individual Hands in the Guardian Manuscripts”

Schreibstil erlaubt Aufschlüsse über

- **Zusammengehörigkeit von Manuskriptfunden**
- **zeitliche Einordnung**
- **Art und Weise der Manuskripterstellung**
- **Bedeutung eines Manuskripts**

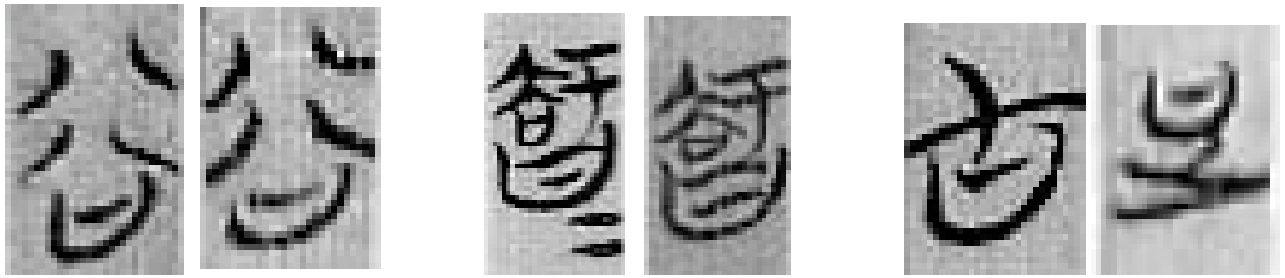
Charakterisierung von Schreibstilen

- **Schrifttyp**
  - regionale und historische Charakterisierung
  - übereinstimmende Zeichenformen und -strukturen
- **Schreibstil**
  Ausführungsart für einen Schrifttyp, zB gruppiert in
  A kunstvoll, regelmäßig, kontrolliert
  A* zwanglos, kursive Form von A
  B schlicht, schmucklos, gerade Striche
  C dynamisch, aufblühend, barock, wellig
  D kunstvoll, ornamental, schlank
- **Schreibermerkmale**
  Genauere Merkmale als bei Schreibstilen, zB
  - Strichqualität
  - morphologische Unterschiede bei einzelnen Zeichen
  - Neigung der Zeichen relativ zur Spalte
  - Winkel zwischen einzelnen Strichen

Schreibermerkmale des Laozi A Manuskriptes (1)

Vergleich der Spalten 5 und 6 (Schreiber B) mit den übrigen Spalten (Schreiber A):

Schreiber A:

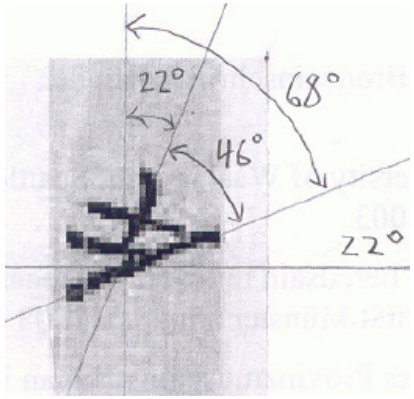


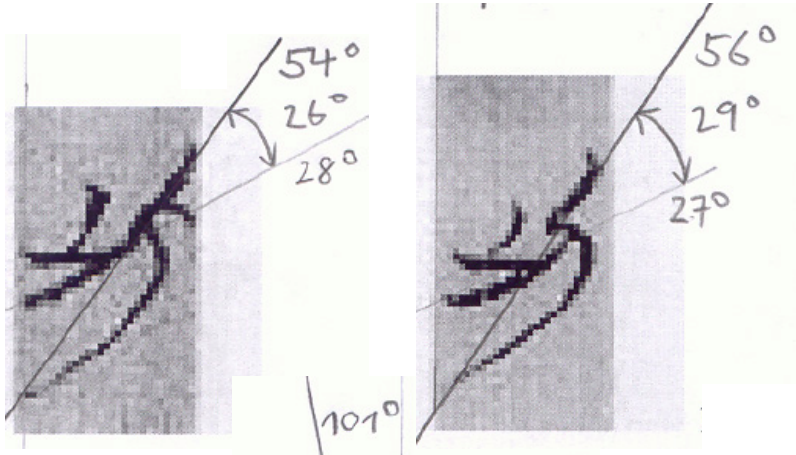
Schreiber B:

Schreibermerkmale des Laozi A Manuskriptes (2)

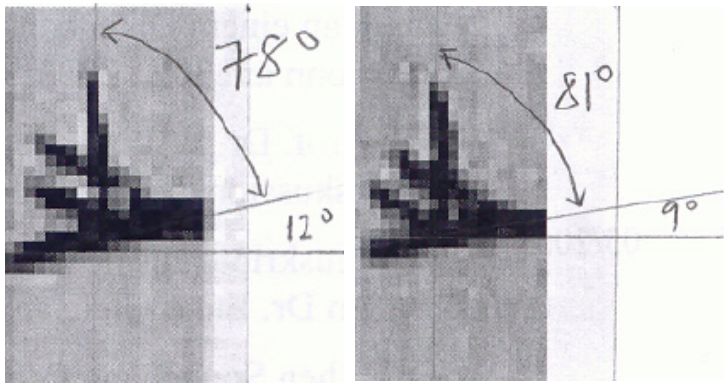
Vergleich der Spalten 5 und 6 (Schreiber B) mit den übrigen Spalten (Schreiber A):

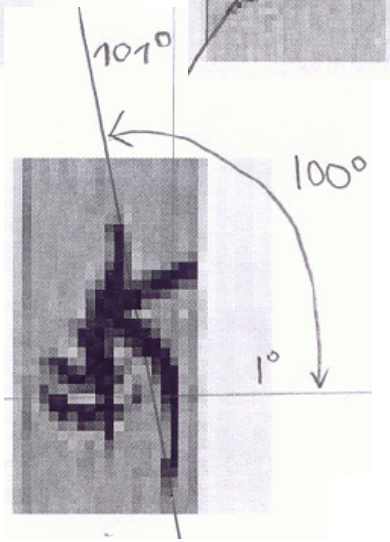
Schreiber A:





Schreiber B:





Was kann (vielleicht) ein Computer ermitteln?

- **Gleiche Zeichen erkennen**
  - in einem Manuskript
  - in verschiedenen Manuskripten
- **Merkmale berechnen**
  - Strichart
  - morphologische Differenen
  - Neigungen relativ zur Spalte
  - Winkel innerhalb eines Zeichens
- **Diskriminierende Merkmale entdecken**
  - "Gesamteindruck" objektivieren
  - neue Merkmale entdecken
- **Schreiberklassifikation**
  - Variabilität bestimmen
  - Statistisch fundierte Entscheidungen

Kleine Einführung in die digitale Bildverarbeitung

Geschichte vom Maschinellen Sehen (1)

Vision für Computer Vision

Selfridge 1955: *" ... eyes and ears for the computer"*

Erste Anwendungen von Bildverbesserung und Bildverarbeitung

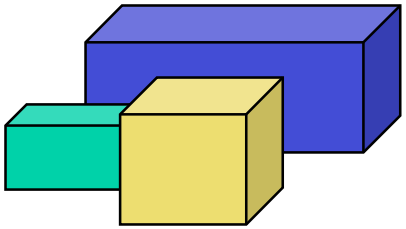
Weltraummissionen, Luftbildauswertung

Zeichenerkennung

=> Mustererkennungsparadigma

Blockswelt, eingeschränkte Domänen

Roberts 1965: 2D => 3D



Natürliche Szenen mit Bewegung, Video

Nagel 79: Digitalisierung und Analyse von Verkehrszenen

Visuelle Agenten

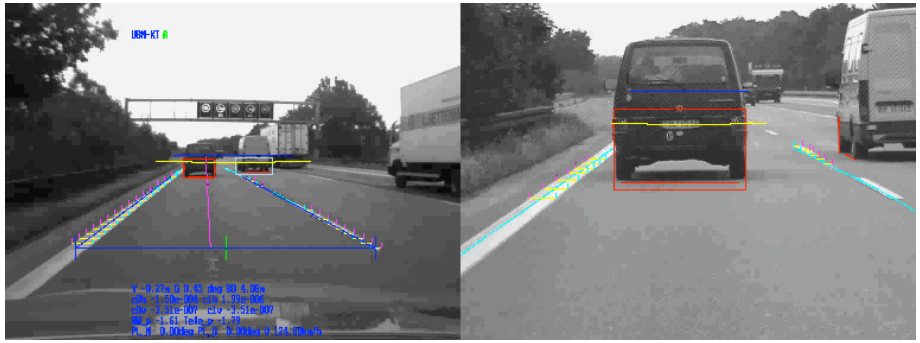
Bajcsy 1988: Active Vision

Geschichte vom Maschinellen Sehen (2)

Visuelle Fahrerassistenzsysteme

Dickmanns 1996:

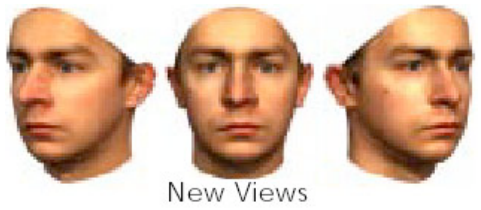
Autonome Navigation auf Autobahnen



Erkennen von Gesichtern

Bülthoff 2002:

Ansichtenbasierte Gesichtsmodelle



Ereigniserkennung und Szeneninterpretation

Hongeng 2003:

Erkennen krimineller Handlungen

Stand der Technik bei Zeichenerkennung und Zeichenanalyse

- **Leistungsstarke kommerzielle Systeme zur Erkennung gedruckter und handgeschriebener Zeichen (OCR) für zahlreiche Schriften**
  - **Fortschritte durch 10^5-fache Steigerung der Rechengeschwindigkeit seit 1960**
  - **Adaptive und lernende Verfahren**

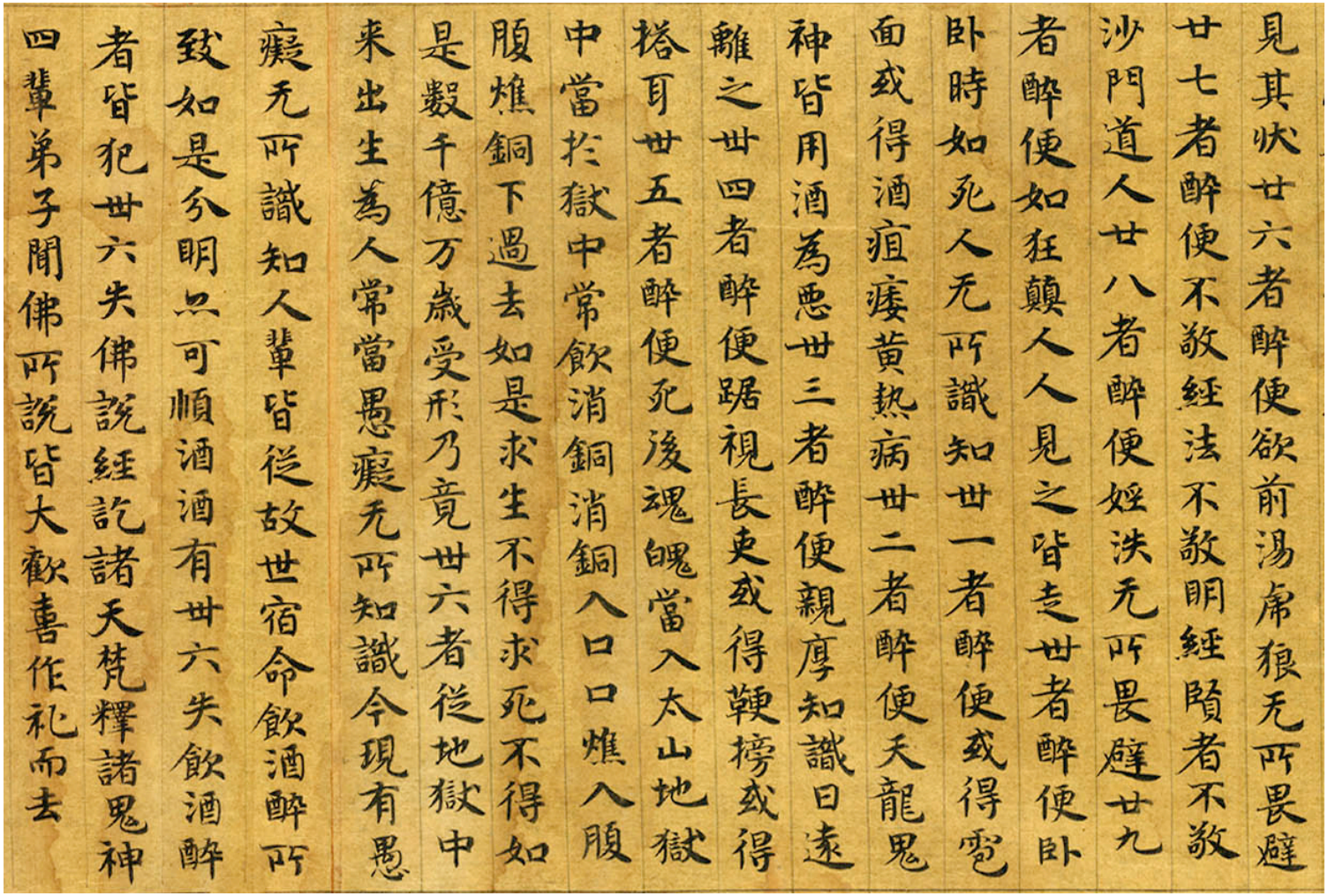
- **Umfassendes Repertoire von Algorithmen zur Ermittlung von Objektmerkmalen und geometrischen Eigenschaften.**

- **Aktuelle Verbesserungen bei Segmentierungsverfahren ermöglichen neue Anwendungen unter schwierigeren Bedingungen.**

Strukturelle Manuskriptanalyse

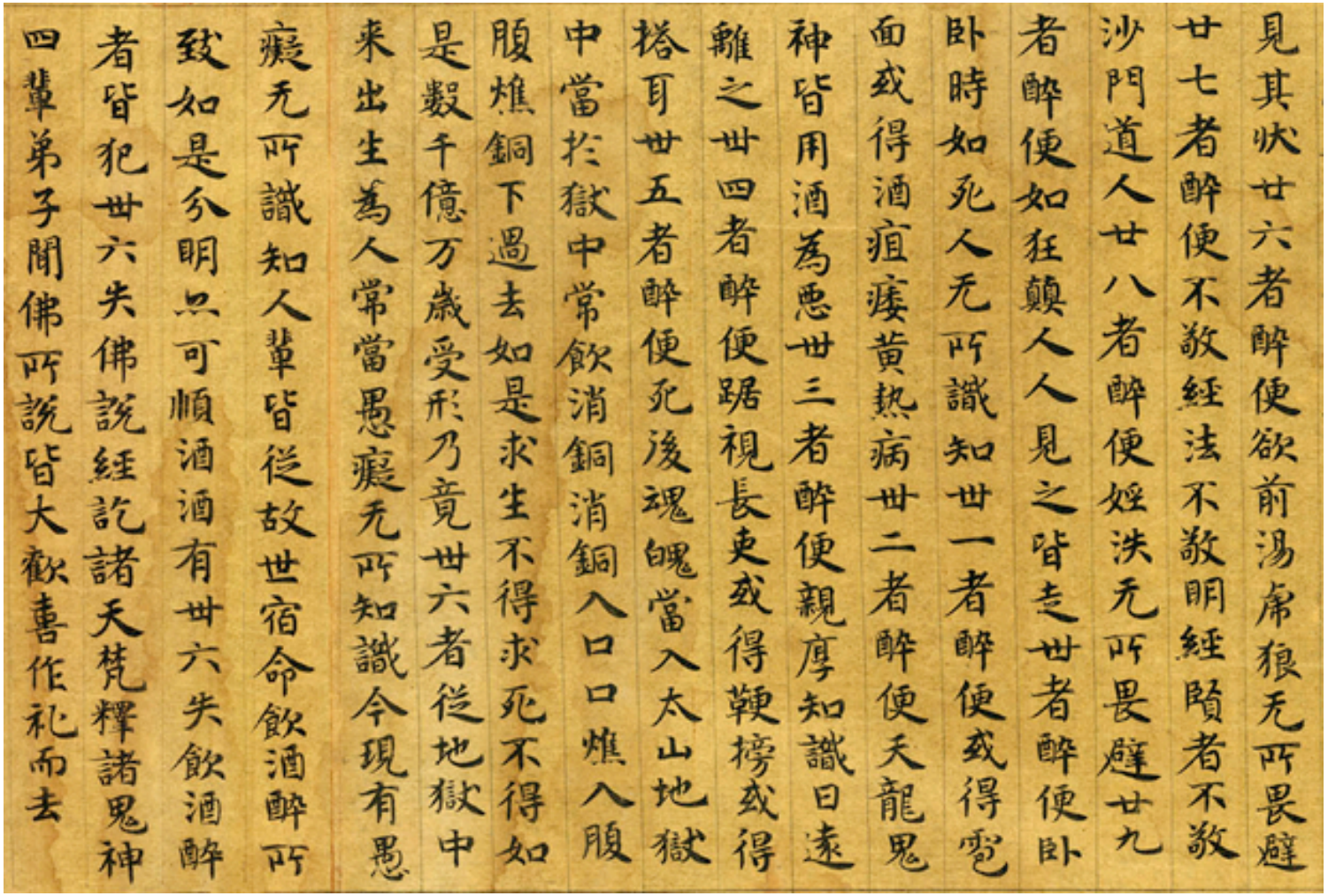
Digitalbild eines Manuskriptes

1236 x 834 Pixel



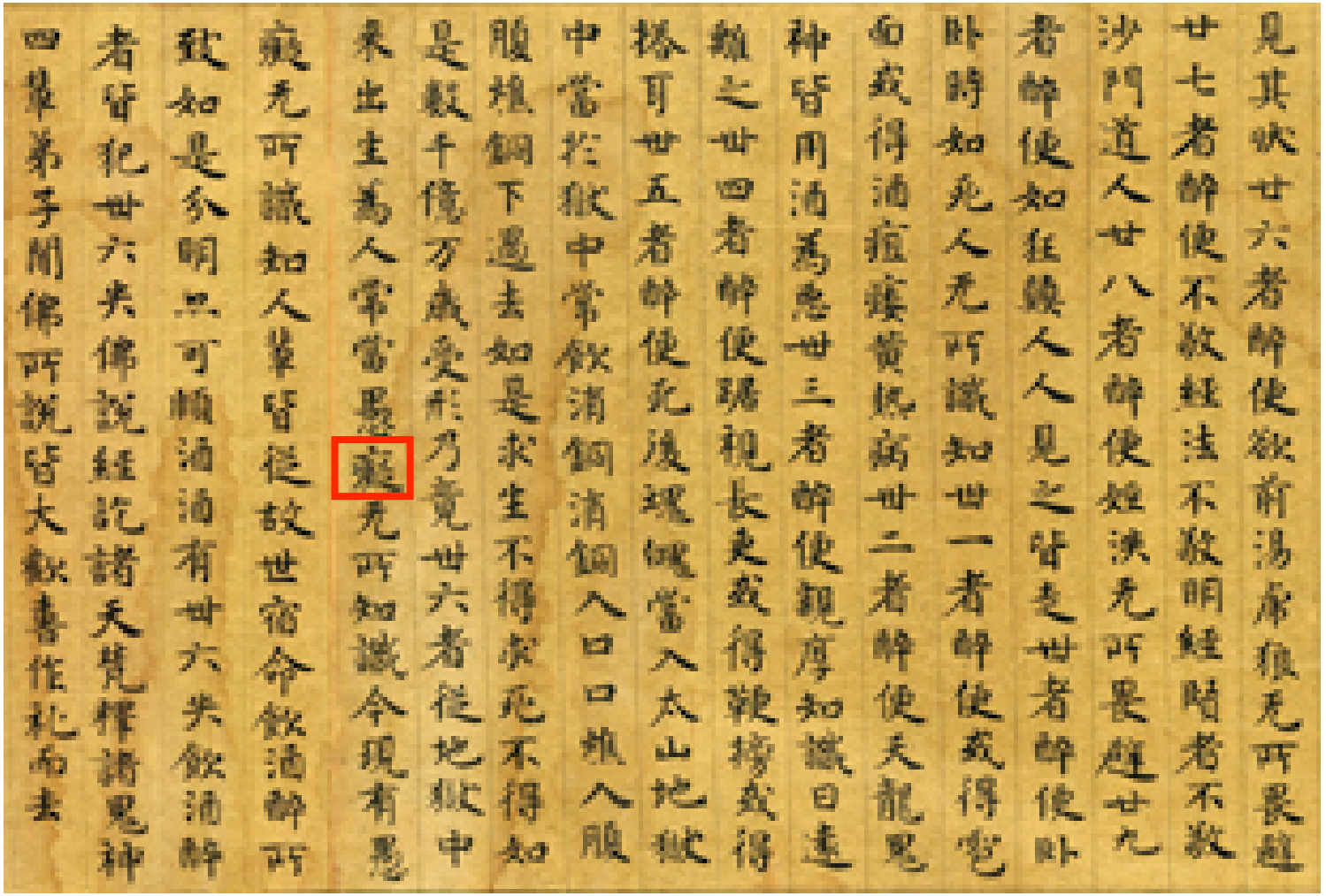
Digitalbild - halbe Auflösung

618 x 417 Pixel



Digitalbild - viertel Auflösung

309 x 208 Pixel



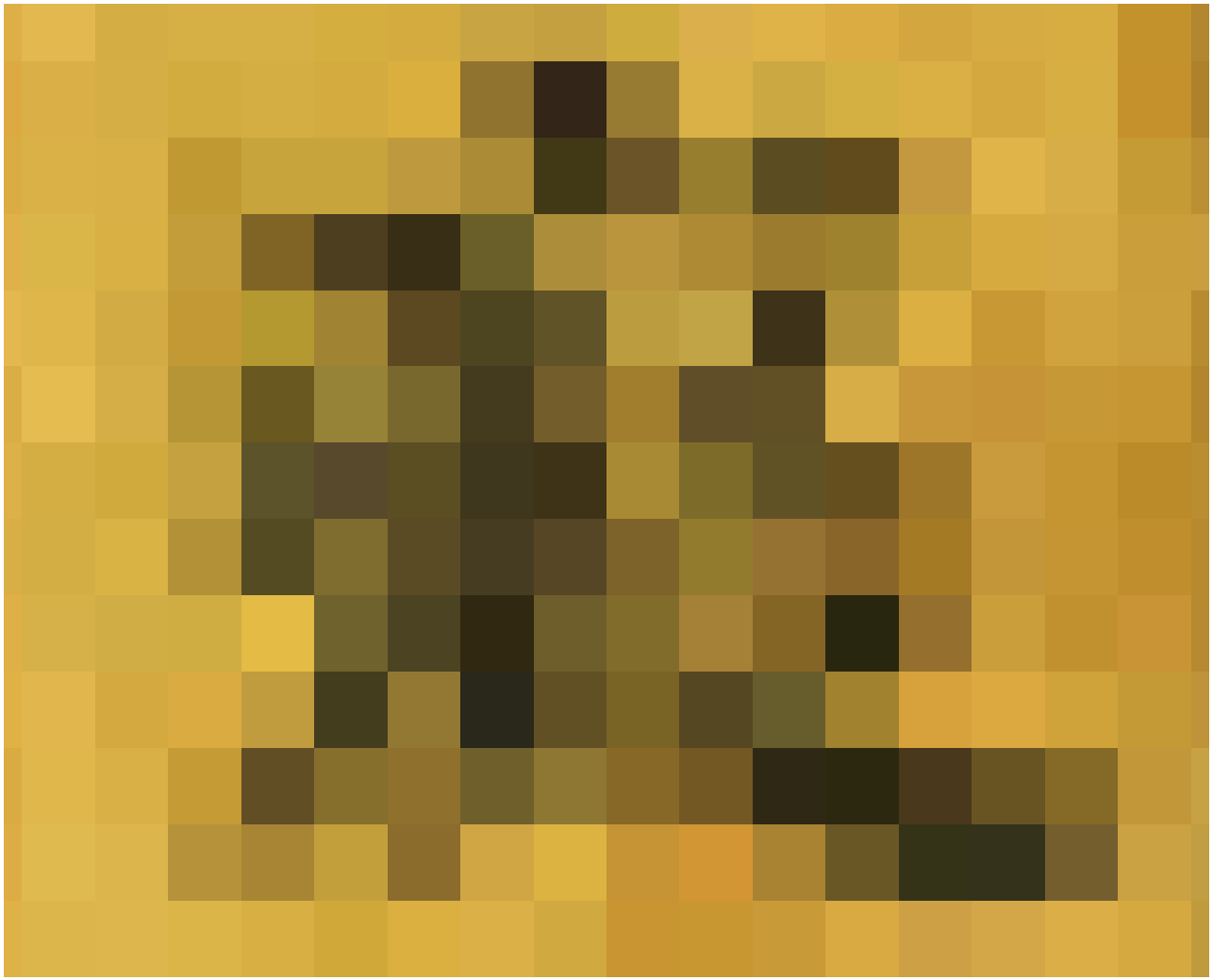
Ausschnitt - volle Auflösung

60 x 50 Pixel

Ausschnitt - halbe Auflösung

30 x 25 Pixel

Ausschnitt - viertel Auflösung



16 x 13 Pixel

Layout-Analyse

Aufgaben der Layout-Analyse

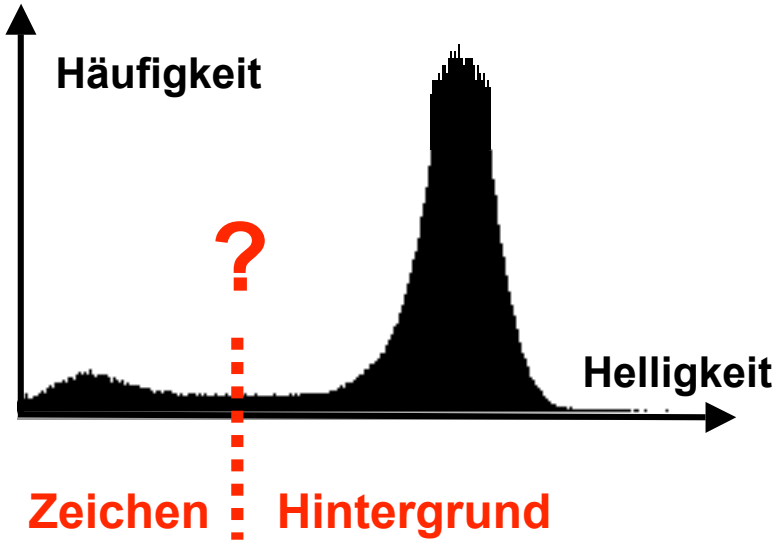
- **Bestimmen von Spalten**
- **Isolieren einzelner Zeichen**
- **Bestimmen von Layout-Parametern**
  - **Spaltenabstände**
  - **Spaltenneigung**
  - **Zeichenabstände**

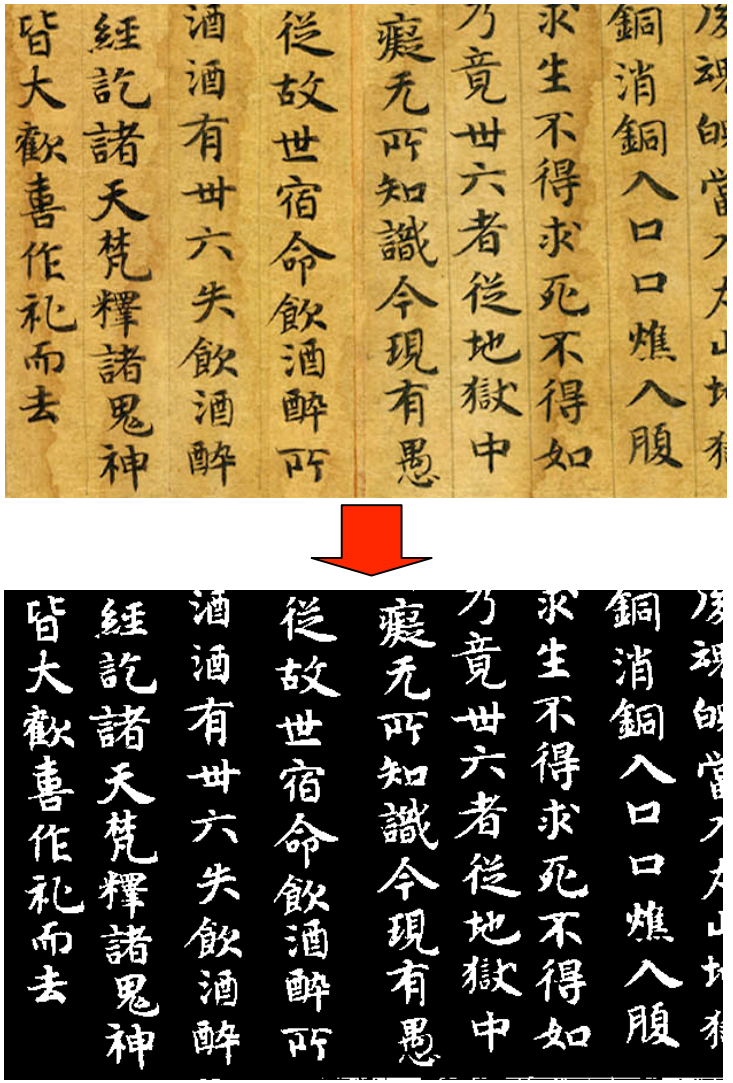
  } **Einzelwerte, Mittelwert, Varianz**
- **Qualitätsmaße, stilistische Merkmale**

Binarisieren

Umwandeln in ein S/W-Bild durch Schwellwertvergleich

Schwellwertbestimmung aus Histogramm:





Feinstruktur unterdrücken

Tiefpassfilterung mit

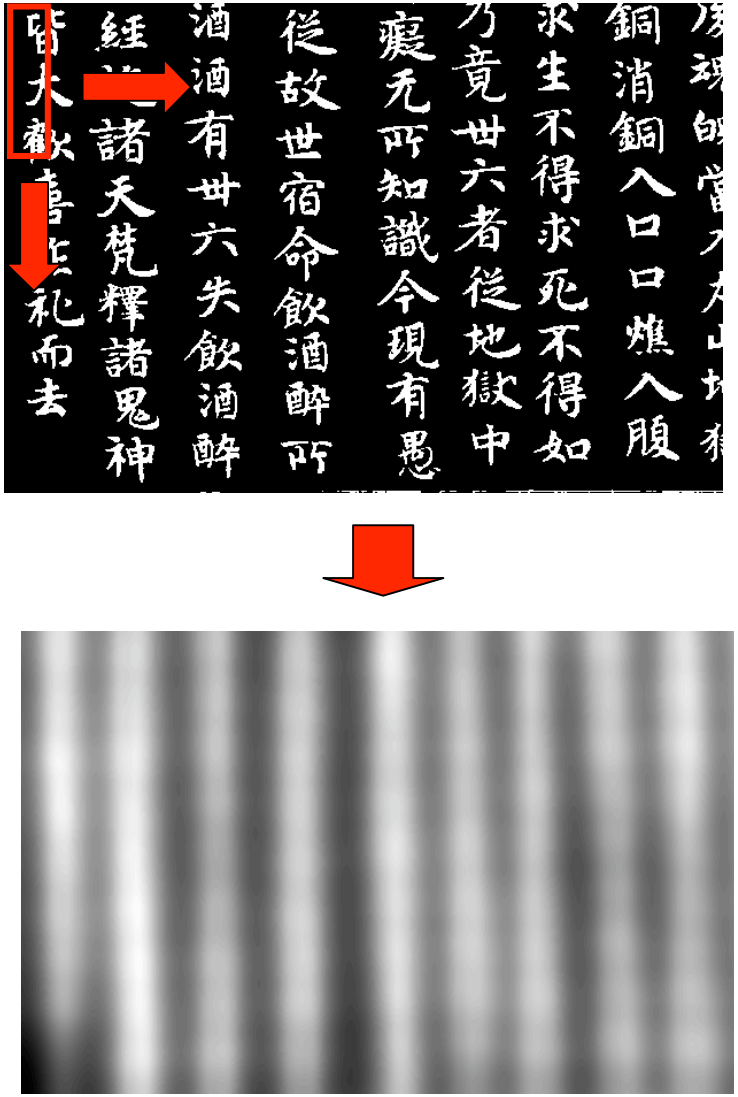
- starker Glättung vertikal

- schwacher Glättung horizontal

"Filtern" im Digitalbild:

Jedes Pixel durch die gewichtete Summe seiner Umgebung ersetzen

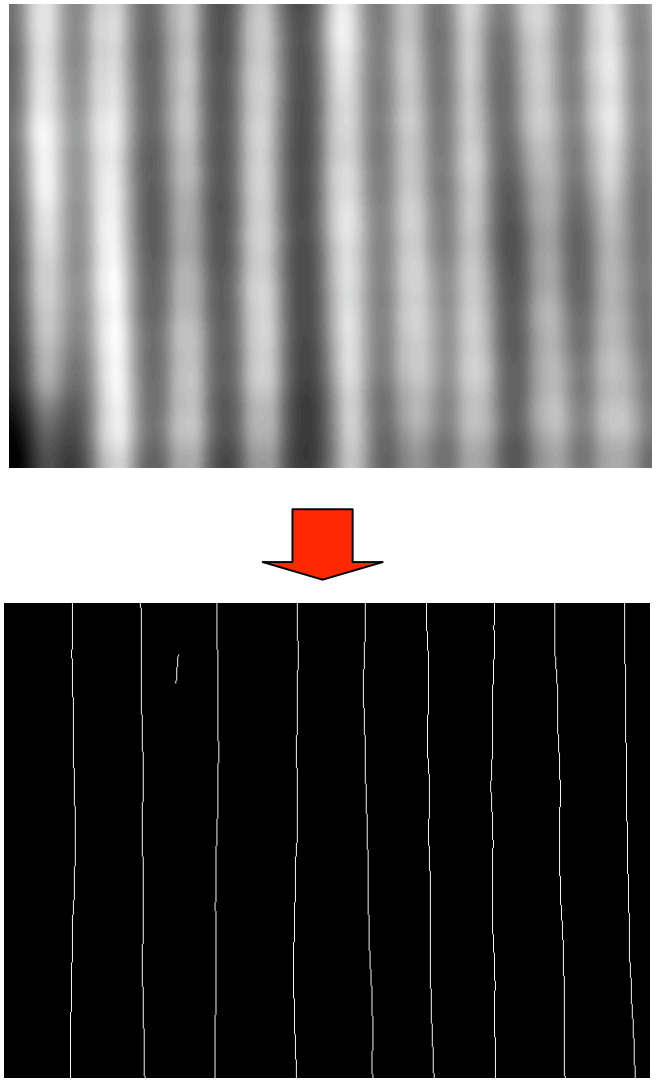
$$g' = \sum w_{ij}\, g_{ij}$$



Spaltengrenzen bestimmen

Die lokalen Minima in jeder Zeile ergeben Anhaltspunkte für Spaltengrenzen.

Falsche Linienstücke können durch Vergleich mit dem überwiegenden Spaltenabstand beseitigt werden.



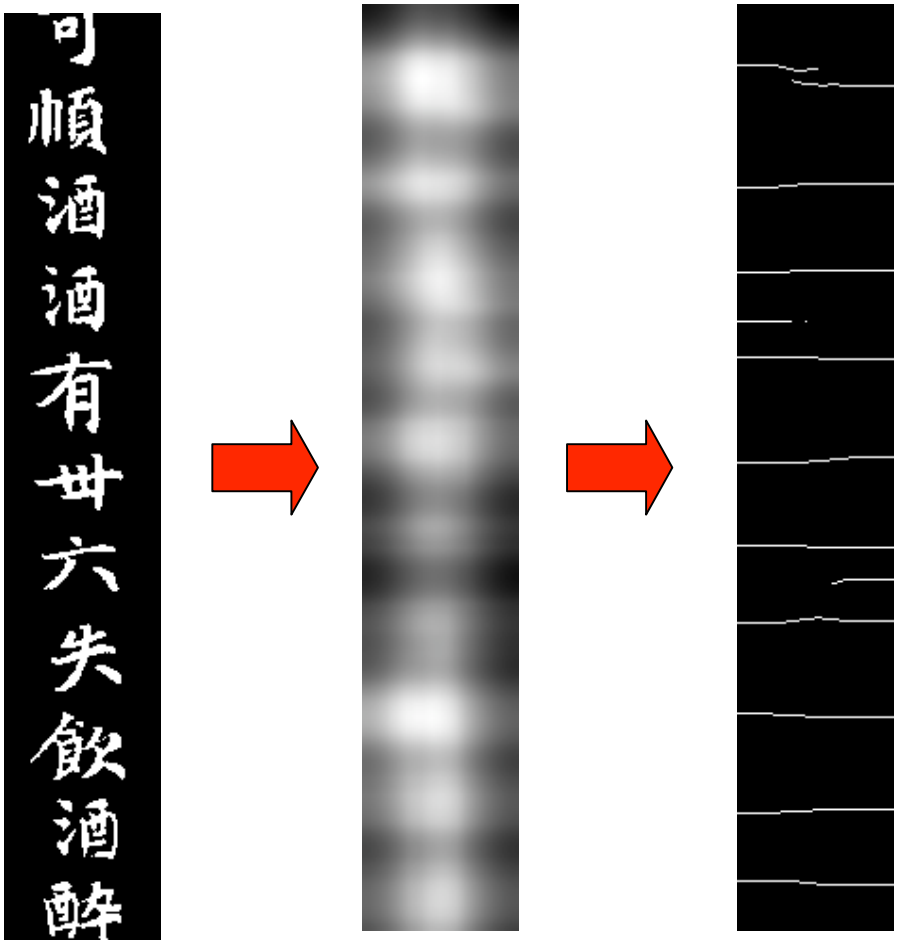
Zeichen isolieren

Verfahren analog zur Spaltenbestimmung:

Tiefpassfilterung mit

- **starker Glättung horizontal**
- **schwacher Glättung vertikal.**

Lokale Minima in jeder Spalte ergeben Trennlinien.



Verbleibende Probleme

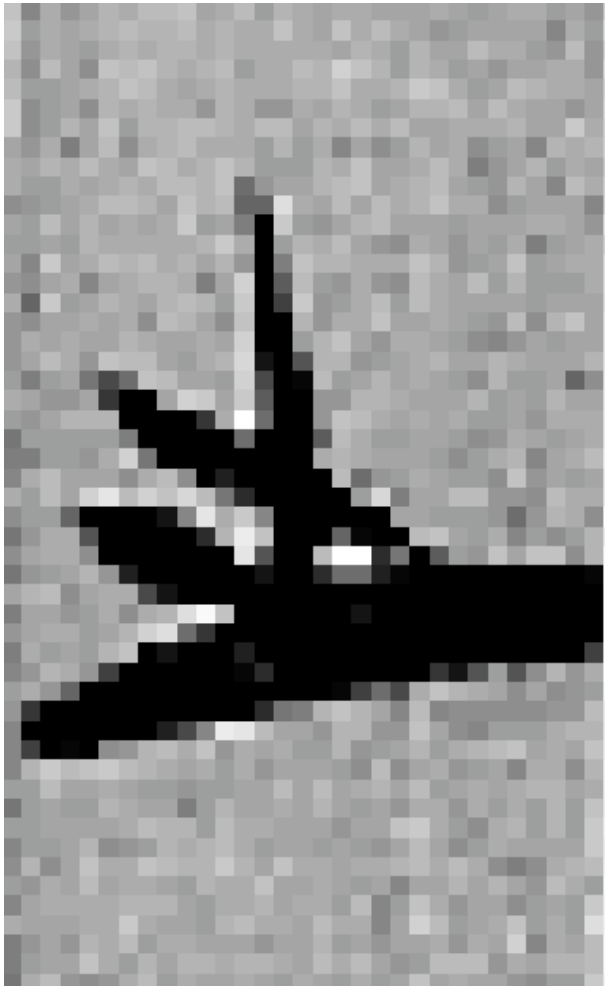
Eine Trennung aufgrund einer festen Filtereinstellung ist nicht zuverlässig.

Filter muss lokal angepasst werden.

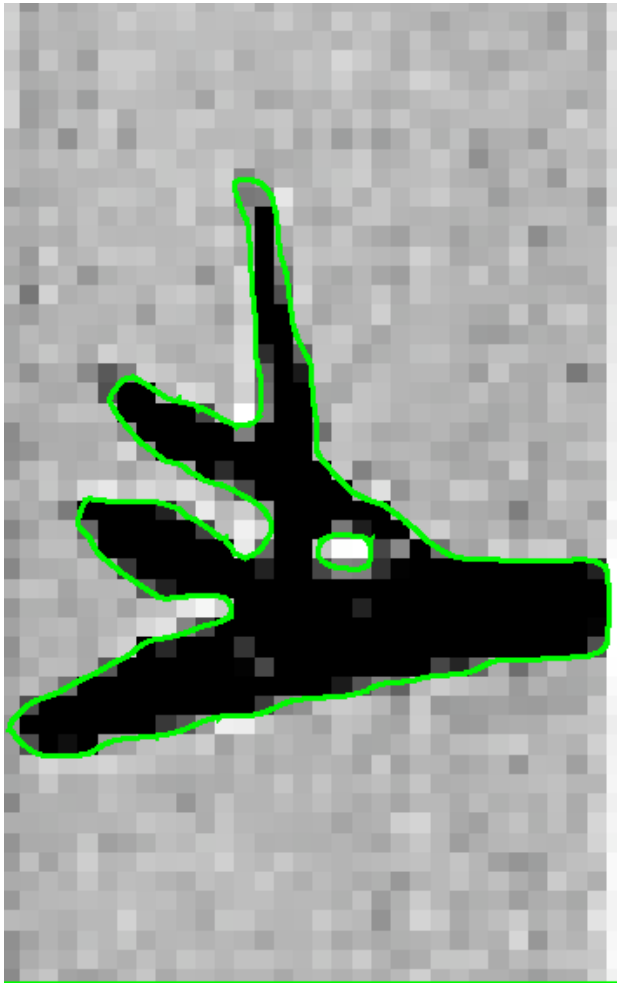
Strukturelle Zeichenanalyse

Segmentierung

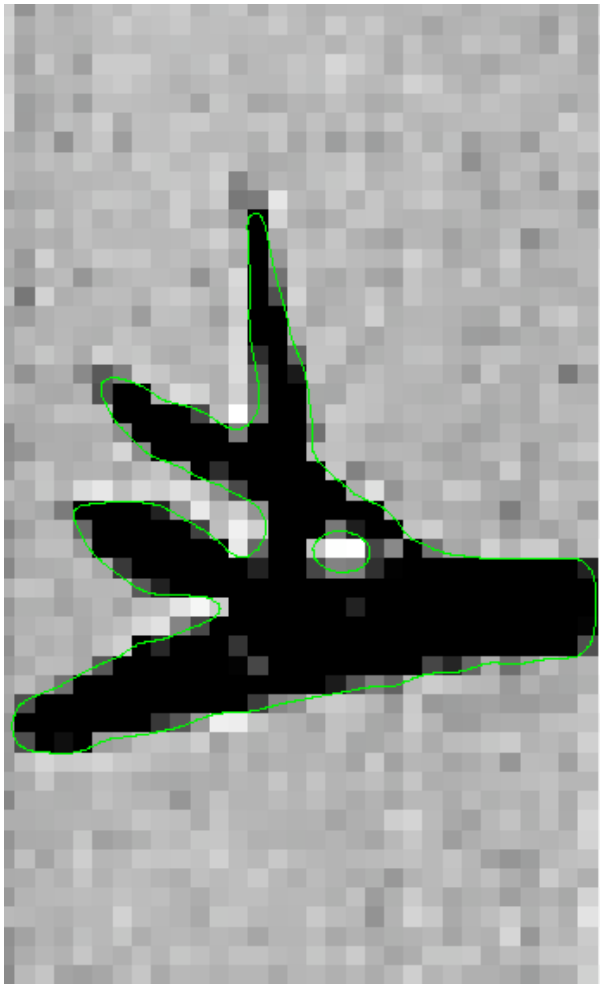
Original



Subpixel-Wasserscheidenverfahren



Subpixel-Schwellwertverfahren



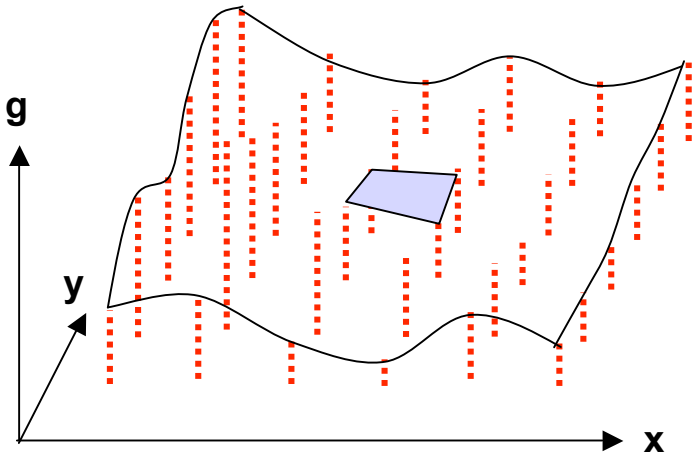
Subpixel-Segmentierungsverfahren

Subpixel-Schwellwertverfahren

A Bestimme Schwellwert aus Histogramm

B Erzeuge kontinuierliches Bild durch Spline-Interpolation zwischen den Pixeln des Originalbildes

C Verfolge Schwellwert-Höhenlinie in kontinuierlichem Bild



Spline-Interpolation 5-ter Ordnung

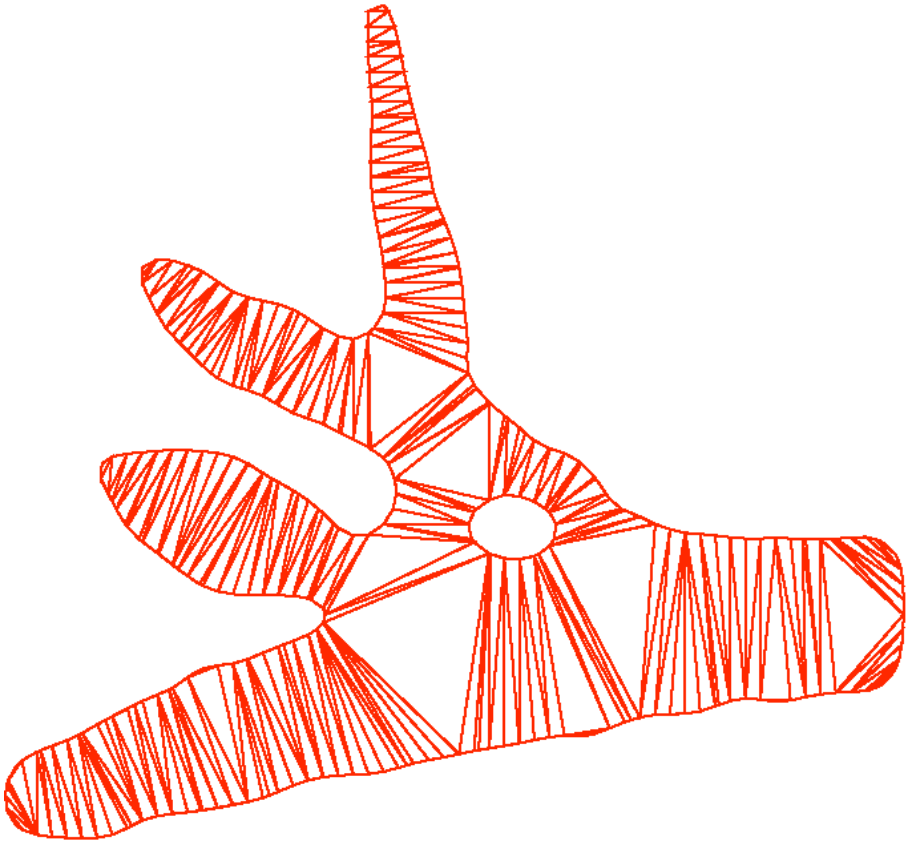
Subpixel-Wasserscheidenverfahren

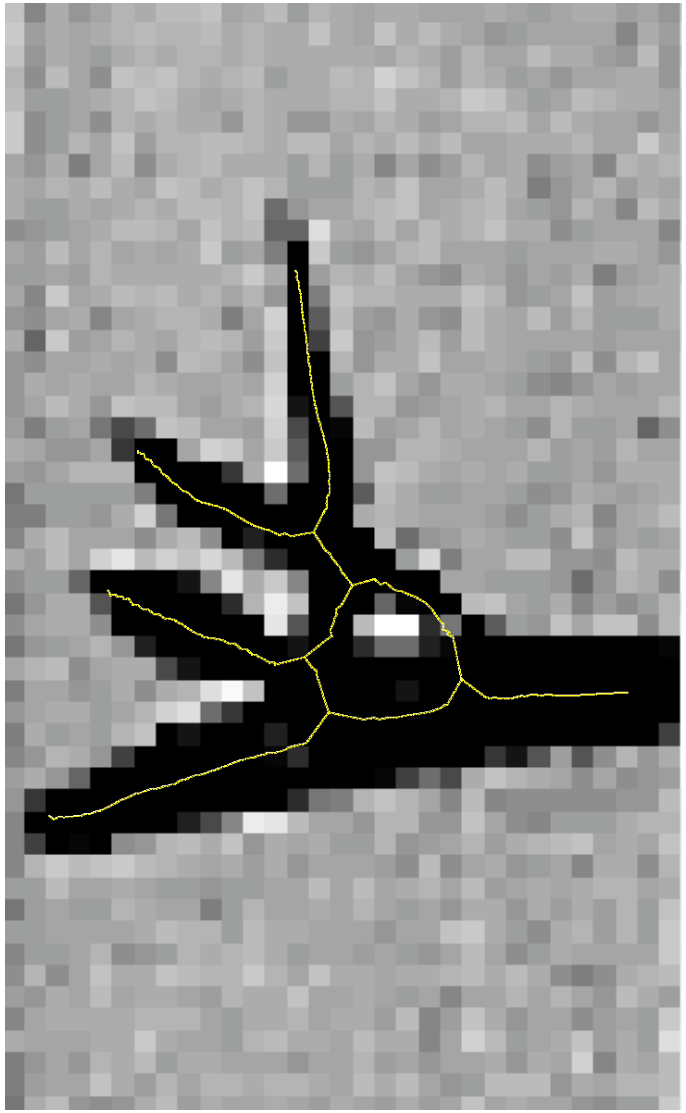
A Erzeuge kontinuierliches Bild durch Spline-Interpolation zwischen den Pixeln des Originalbildes

B Erzeuge Gradientenbild durch Differenzierung des (analytischen) interpolierten Bildes

C Verfolge Maxima im Gradientenbild

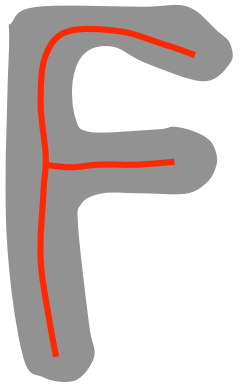
Skelettierung



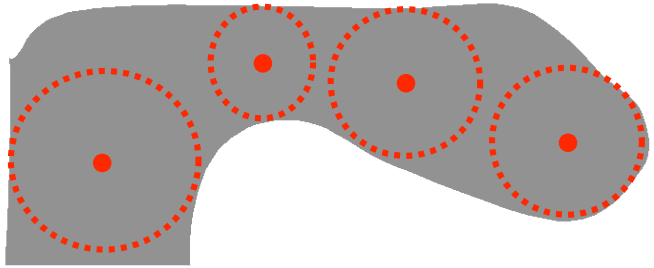


Medialachsen-Transformation

Bestimme Linienstruktur ("Skelett"), die die wesentliche Form eines Zeichens wiedergibt.



Die Medialachsen-Transformation (MAT) besteht aus allen Pixeln, die mehr als einen nächsten Randpunkt haben.



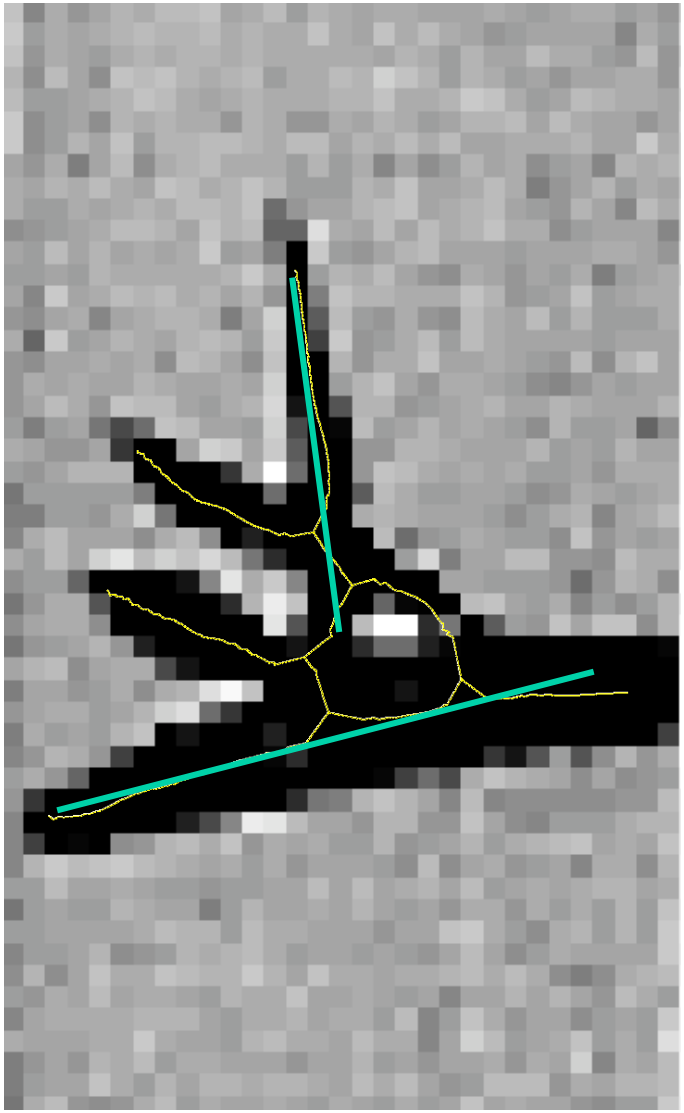
Berechnung der MAT z.B. durch den "Grasfeueralgorithmus":

Entferne Pixel gleichmäßig vom Rand wie eine Flammenfront. Wo sich zwei oder mehr Fronten begegnen, ist die Medialachse.

Orientierungsbestimmung

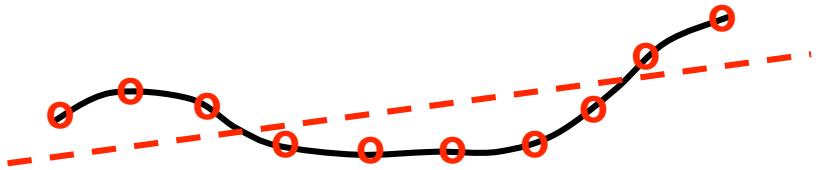
Anpassen einer Geraden an einzelne Linienzüge des Skeletts.

Hier: Manuelle Auswahl der Linienzüge



Geradenanpassung

Welche Gerade gibt die Richtung einer Linie am besten wieder?



A Repräsentiere die Linie näherungsweise durch Punkte

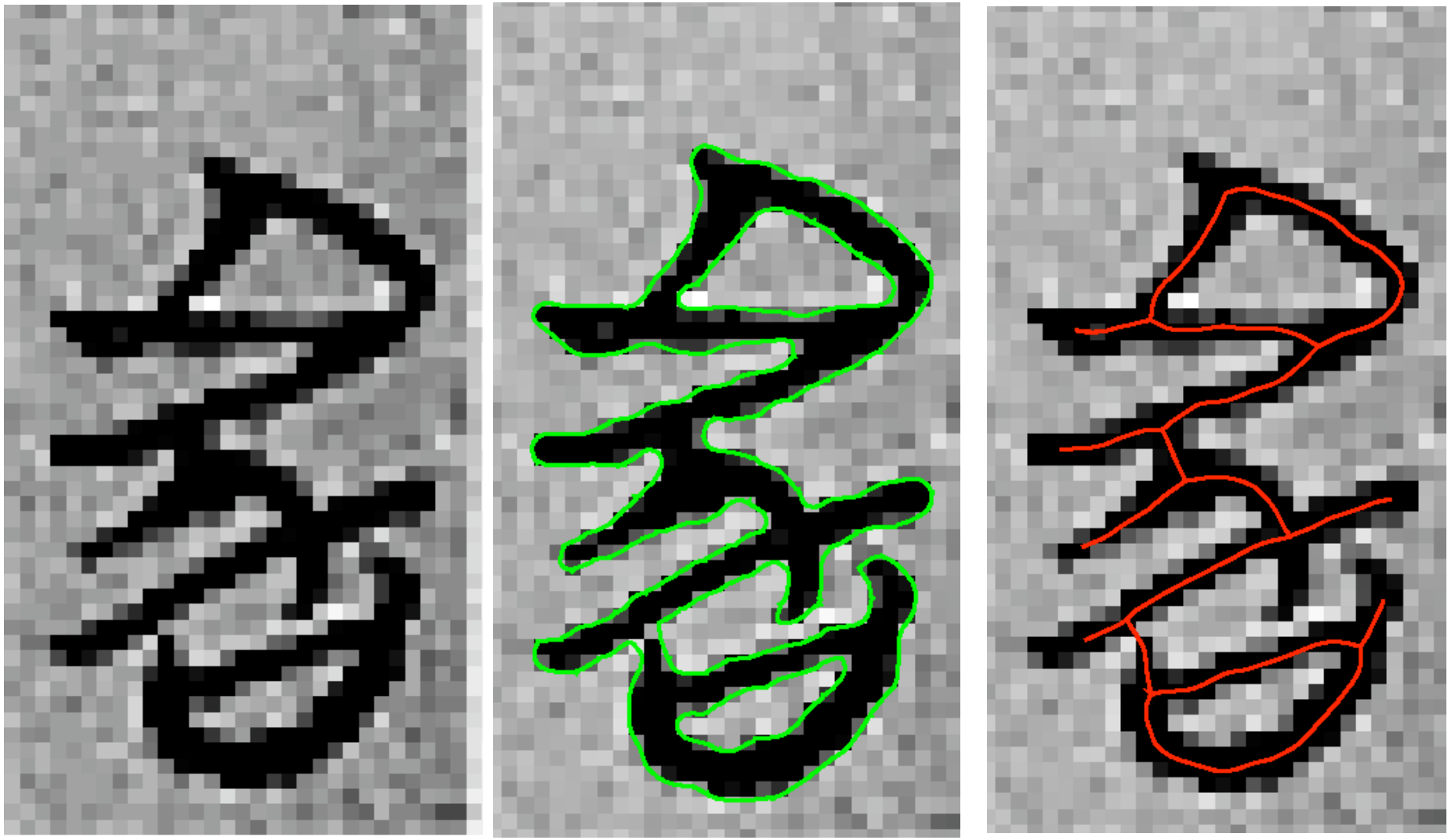
B Bestimme die Gerade, zu der die Punkte insgesamt den geringsten Abstand haben:

- Bestimme Mittelwert der Punkte $[m_x\ m_y]$, Gerade geht durch $[m_x\ m_y]$
- Bestimme Streumatrix

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} = \begin{bmatrix} \Sigma\,(x_i-m_x)^2 & \Sigma\,(x_i-m_x)(y_i-m_y) \\ \Sigma\,(x_i-m_x)(y_i-m_y) & \Sigma\,(y_i-m_y)^2 \end{bmatrix}$$

C Richtung des größten Eigenvektors von S bestimmt Gerade

Zeichenstruktur - Beispiel 2



Clustering und Datamining

Clustering

Gruppieren von Objekten aufgrund eines Ähnlichkeitsmaßes

- **gewichteter Euklidischer Abstand für numerische Werte**
  z.B. Unterschied eines Längenverhältnisses, eines Winkels
- **feste Abstandswerte für symbolische Kategorien**
  z.B. Unterschied eines Schreibstils

Zahlreiche Clustering-Verfahren stehen in kommerziellen Programmsystemen zur Verfügung.

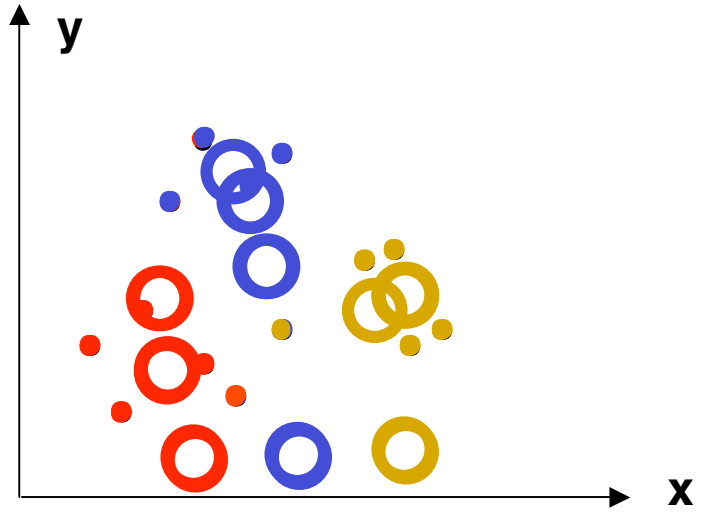
Vorsicht: Ergebnis hängt von Gewichtung der Unterschiede ab!

K-means Clustering

Erfordert Vorgabe der Cluster-Anzahl

A	**Initialisiere Clusterzentren**
B	**Ordne Ereignisse nächstliegenden Clusterzentren zu**
C	**Neue Clusterzentren sind Schwerpunkte der zugeordneten Ereignisse**
D	**Wiederhole B und C, bis keine Veränderungen mehr auftreten**

Beispiel:
Gruppiere Manuskripte entsprechend zweier Merkmale x und y.



Was ist Datamining und Knowledge Discovery (DMKD)?

Klassische Statistik

Maschinelles Lernen

Mustererkennung

} **Data Mining und Knowledge Discovery wird Teilgebiet der KI (1990)**

Identifikation gültiger, neuer, potentiell nützlicher und letztendlich verständlicher Muster in großen Datenbeständen durch nicht-triviale Prozesse

Data Mining:
Analyse, Anbieten von Hypothesen

Knowledge Discovery (Wissensentdeckung):
Bewerten und Deuten

Datamining für die Manuskriptanalyse

Entdeckung von visuellen Merkmalen für die Bestimmung von Gemeinsamkeiten zwischen Manuskriptfundstücken

Potentieller Aufschluss über

- **gemeinsame Herkunftsregion**
- **gemeinsamen Schrifttyp**
- **gemeinsamen Schreibstil**
- **gemeinsamen Schreiber**

These:
Rechner können Merkmale und Merkmalsassoziationen entdecken, die von Menschen nicht wahrgenommen oder übersehen werden.

Assoziationsregeln entdecken

Datamining bietet ausgereifte und effiziente Verfahren, z.B. zur Warenkorbanalyse:

"Kunden, die Bier und Pizza einkaufen, kaufen auch gerne Kartoffelchips"

Übertragen auf Manuskriptanalyse:

"Schreiber X hat in Zeichen Y häufig einen kleineren Aufstrichwinkel als andere Schreiber"

"Es gibt Manuskripte, in denen der Querstrich in Zeichen X meist kürzer ist als anderswo, das Strichende in Zeichen Y1, Y2 und Y3 meist spitzer ausläuft als normal, und die Zeichenhöhe ungewöhnlich konstant ist."

Realisierung von Datamining auf Manuskripten

- **Manuskriptsammlung zusammenstellen und digitalisieren**
- **Merkmalsrepertoire aufbauen und Berechnungsverfahren implementieren**
- **A-PRIORI-Algorithmus sucht Assoziationen mit aufsteigender Anzahl von Elementen**
- **Steuerung durch Vorgabe von Support und Konfidenz**

Thema für ein geplantes Forschungsprojekt!

Zusammenfassung

- **Manuskriptforschung kann von der Analyse visueller Merkmale profitieren.**
- **Digitale Bildverarbeitung bietet ein reiches Methodenrepertoire für Layout-Analyse und strukturelle Zeichenanalyse.**
- **Neuere Forschungsergebnisse aus der Bildverarbeitung erlauben Anwendungen für schwierigere Bilddaten.**
- **Bei der Auswertung einer möglicherweise unübersichtlich großen Zahl von visuellen Merkmalen können Clustering- und Datamining-Verfahren helfen.**