

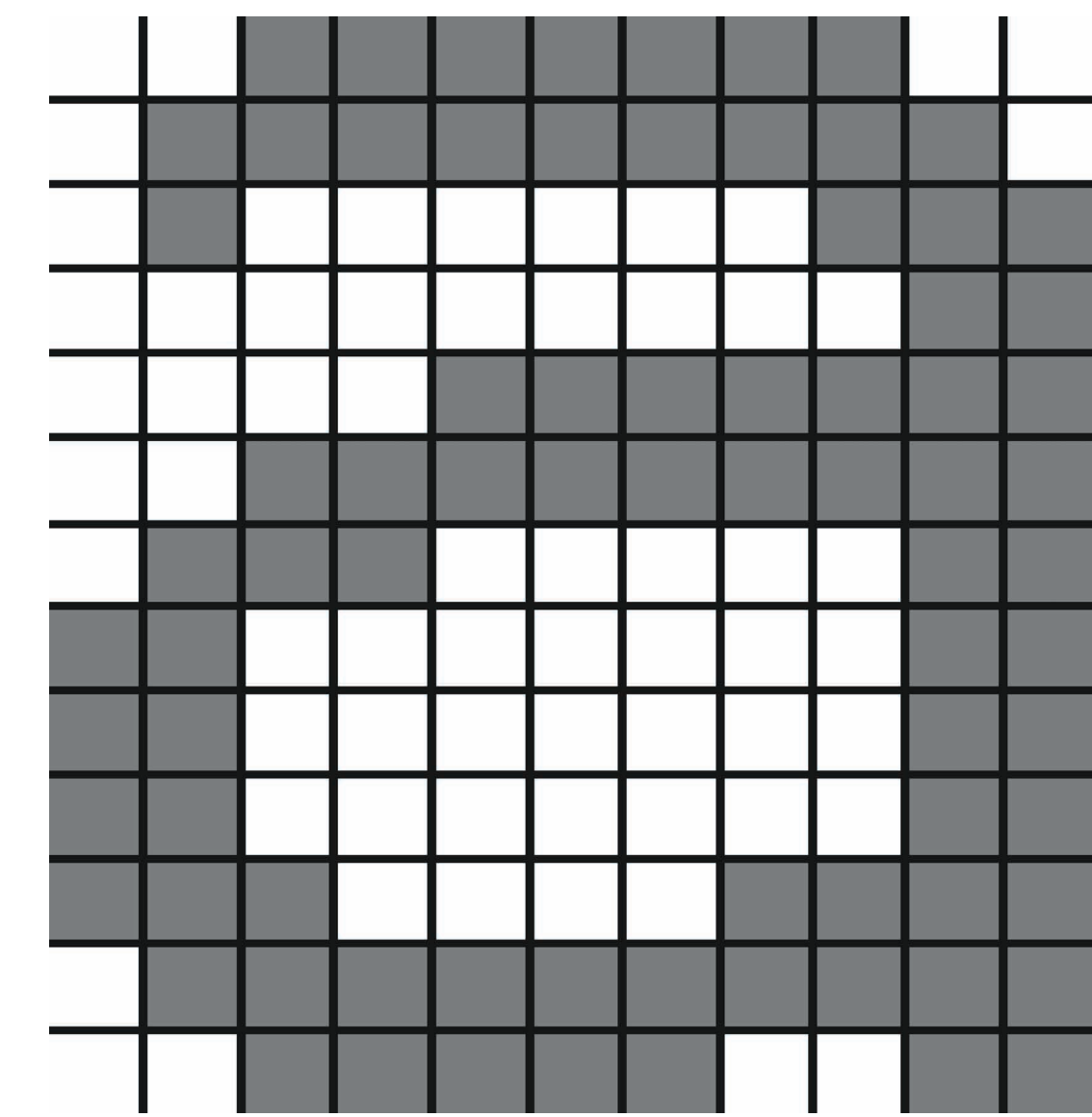
Texterkennung



Problem

Bei der Texterkennung geht es um das Herausfiltern des Textes aus gescannten Buchseiten. Die Seite liegt als Bilddatei vor, welche aus vielen kleinen Punkten (Pixeln) besteht (s. Abb.). Diese Pixel sind hier entweder weiss oder schwarz.

An OCR engine filters the text of a scanned book page. The pages are on hand as image files which are made up of pixels (cf. fig.). Those pixels are either black or white.



Idee/Idea

Das Problem der Erkennung eines ganzen Textes wird aufgeteilt in die Erkennung von einzelnen Zeichen. Zur Isolierung eines Zeichens wird zuerst entlang aller Bildzeilen jedes schwarze Pixel gezählt. Die Anzahl dieser Pixel wird am Zeilenende als Kurvenausschlag dargestellt (s. Abb.). Kein Ausschlag signalisiert einen Zeilenzwischenraum. Der Text kann nun in Zeilen zerlegt werden. Dasselbe wird nun ebenfalls in vertikaler Richtung für jede einzelne Zeile vorgenommen. Die so freigestellten Zeichen werden mit einer Datenbank verglichen und das Zeichen mit der kleinsten Abweichung wird eingesetzt. Da bei gewissen Zeichen eine grosse Ähnlichkeit besteht, können mit einer Wortprüfung Fehler verhindert werden.

Texterkennung
ist gar nicht
so schwer...



Texterkennung
T W Z P R I N T

The problem of recognising a whole text is split into recognising individual characters. Along each pixel row the black pixels are counted to isolate the text lines: No black pixel along a row indicates a space between two lines. With the same procedure along the verticale, the characters can be separated. Afterwards the OCR engine compares each single character with a database and chooses the one with the lowest difference. Some characters look quite similar, therefore a word check is able to prevent wrong choices.