3

OCR Technologies for Machine Printed and Hand Printed Japanese Text

Fumitaka Kimura

3.1 Introduction

Commercial products of Japanese OCR are classified into form processing OCR and document OCR. The form processing OCR is mainly aimed at reading handwritten characters filled in a printed form with blank spaces for information. Its implementation is either by hardware device or computer software. Meanwhile the document OCR is mainly aimed at reading machine printed documents such as newspapers, magazines and general documents. Its implementation is by computer software in most cases.

The OCR products are used to save labour and time of keyboard entry in various business tasks including customer support, sales management, financial account, questionnaire survey, etc. The OCR software is also used as built-in OCR applications in word processing, spread sheet processing, full text retrieval, facsimile OCR, filing system and document workstation.

Many technologies relating to image processing, pattern recognition and linguistic processing are employed in Japanese OCR systems. This chapter deals with the OCR technologies for pre-processing, feature extraction, classification, dimension reduction and learning as well as the performance evaluation of those techniques.

3.2 Pre-Processing

Pre-processing for Japanese text recognition includes text line segmentation, character segmentation and relating normalization techniques for skewed documents, slant of characters and size of characters.

3.2.1 Text Line Segmentation and Skew Correction

In text line segmentation, horizontal projection of a document image is most commonly employed, when actual text line orientation is horizontal

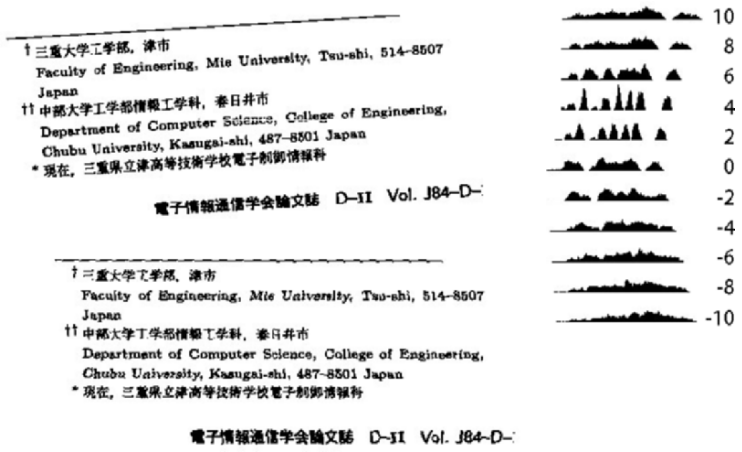

Fig. 3.1. Example of skew correction.

(Figure 3.1). If lines are well separated and are not skewed, the horizontal projection has well separated peaks and valleys. These valleys are easily detected and used to determine the location of boundaries between lines. Main difficulties of this simple strategy are encountered when dealing with skewed lines. The peaks and valleys are not distinctive for skewed document images, and the text lines are not separable by horizontal boundary lines. A typical approach to handle the problem of skew is to estimate and correct the skew preceding the line segmentation.

Crossing Point Method [5]

Skew correction is generally performed in two steps. The skew is estimated in the first step and then a rotation transformation is applied to correct the skew. The basic strategy of the skew estimation is to find the direction in which the projection of the document has maximum separability regarding the peaks and valleys. In the crossing point method, only crossing points are counted to obtain the projection. Where the crossing point is a pixel with value "1" adjacent to its left pixel with value "0". The use of the crossing points rather than entire foreground pixels is advantageous both in improving the separability of the projection and in saving computation time. As a simple measure of the separability, variance of the number of crossing points is used. To find the direction that maximizes the separability measure, multiple projections in slightly different direction by one or two degrees are calculated within the range of expected skew.

It is worth observing that this straightforward enumerative search for maximum separability is more efficient than expected, if it is implemented carefully: all the multiple projections are calculated in a single raster scan. Only the crossing points are projected in multiple directions. The mapping