AUTOMATIC RECOGNITION OF PRINTED FARSI TEXTS

BEHROOZ PARHAMI and MAHMOOD TARAGHI
Tehran University of Technology
Tehran, Iran

Index Terms : Character recognition, Computer input, Document input, Farsi, Feature selection, Optical character recognition, Pattern recognition, Persian, Printed text recognition.

SUMMARY

The automatic recognition of printed Farsi (Persian) texts is complicated by several properties of the Farsi script: (a) connectivity of symbols, (b) similarity of groups of symbols, (c) highly variable widths, (d) subword overlap, and (e) line overlap. In this paper, a technique for the automatic recognition of printed Farsi texts is presented and its steps are discussed as follows: (1) digitization, (2) editing, (3) line separation, (4) subword separation, (5) symbol separation, (6) recognition, and (7) post-processing. The most notable contribution of this work is in the algorithms for Steps (5) and (6) above.

In order to understand the algorithms used, it is necessary to introduce a fundamental property of the Farsi script. Farsi font design is done by using a rectangular-tip pen having a length much greater than its width. As the designer moves the pen at certain angles to generate each symbol, lines with varying thicknesses appear. At the unique connection point of two adjacent symbols, the pen moves horizontally on the "connection axis" and produces a line with (maximum) script thickness. The above property along with the fact

that there is no symbol overlap at the connection point is used in the design of the symbol separation algorithm.

The recognition procedure is based on certain geometric properties such as relative width, existence of concavities and loops. In all, twenty geometric features are used for obtaining a 24-bit vector for each symbol to be recognized. The vector thus obtained is matched against templates for the Farsi symbols. In the rare cases when an exact match is not found, the algorithm looks for a best match in which the more reliable features are examined first. The implemented system is capable of storing new templates for each symbol, as they are encountered, in order to improve its performance by learning.

Experience has shown that the minimum acceptable resolution for the digitization hardware is four samples per script thickness, since otherwise many symbols become indistinguishable. This translates into about 100 samples per cm for Farsi type fonts of typewriters and most textbooks. Practical application of the technique to Farsi newspaper headlines has been 100% successful. However, smaller type fonts, which could not be handled by the coarse digitization hardware used, will no doubt result in less than perfect recognition due to the higher effect of digitization noise.

The technique presented here is also applicable with little or no modification to printed Arabic and Urdu texts which use the same alphabet as Farsi. The insight gained from this study will be helpful in the design of an OCR type font for Farsi, Arabic and Urdu languages.