General and Nimble OCR Open System and the Flexible Segmentation Recognition Algorithms

Du Jiangchuan Liu Hongjian

Chongqing University, China

Abstract

This paper presents a general and nimble optical handwritten numerals recognition system which uses G3 fascimile transceiver as the input device. By connecting a G3 fascimile with a special designed board and with a central computer , the system can complets the recognition program very well and it also can locate the recognition program in RAM and/or ROM.

Flexible recognition algorithms can perform raster scanning of the Huffman code produce by the fax only once . The hanger-chain algorithms can seperate handwritten numerals and obtain the first features at the same time. The speciality is flexible . fast. no demand for big memory. The flexible recognition algorithms can read statistical charts of various sizes and formats written by different writers.

#. The structure of system

Based on the market investigations both at home and abroad .more and more government departments at different levels and various enterprises have now used computers widely to collect and process data and there is a great demand for better character recognition system. According to the experience of using past special and dull OCR readers, we have worked out a simple and practical open OCR system which can renew itself according to rapid development of hardware devices. (see fig. 1.)

With the rapride development of the ISI technology . the hardwarw price of OCR machine can be deduced by using advance CPU and advance system struct . And this can made the recognition system more easy to be use . more presise . more flexible and opening . Eespecialy by inserting a special board into the computer to consist the recognition system.

As the properities to price of fax keep growing . we just use Japanesefax Panasonic UF-915 as the input file scanner . As there isn't any changes in the hardware of the fax. the system can be connected with any developing G3/G4 fax with RS-232 interface which is used as synchroinism data communication by means of the HDLC recommandation and the SIO chip . The communication recommandation is written by the CCITT T. 30 standers . and it is located in the ROM of the special board designed

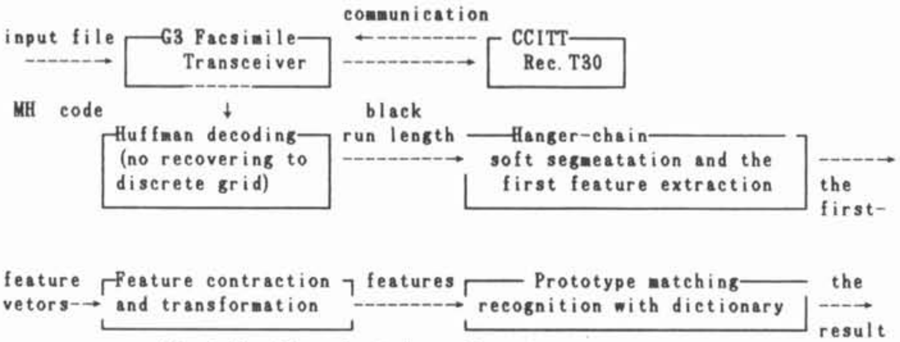
input file
G3 Facsimile Transceiver
communication
CCITT Rec. T30
MH code
Huffman decoding (no recovering to discrete grid)
black run length
Hanger-chain soft segmeatation and the first feature extraction
the first-
feature vetors
Feature contraction and transformation
features
Prototype matching recognition with dictionary
the result

Fig. 1 The flow-chart draw of system

by us. The board can insert into the IBM XT/AT . SO the struct of the system is simple and the fax can do its work dependly and also it can translate the distant data to the center computer by cables or special lines in time . Another advantage of using FAX as input device is that the FAX can put any kind of binary graphs into the computer . and so by changing the program we can do other kinds processing with the data such as editing . replaying etc. not just recognition.

The paper input formats are in orangered. while it passes through the special designed red fluorescent lamp of the FAX . the red lines will all be filtred out. The tool for writing can be pen. pencil. ball-pen. etc.

#. The construcuture of software

The software is devided into four models.

A. the decoding of MH.

B. software segmentation and the first feature extraction.

C. feature abstraction and encoding.

D. dictionary matching and the output results.

1. AS the output of fax is Huffman code. a kind of compacted code . we must decode the MH/MR code to coordinate data first . The decoding algorithms are using the data drive program . use the multiple tree struct and the bit shift technology to realize the higt speed decoding . After decoding we get the coorinate of BRLs. not the discrete grid. So the space is saved. The high speed decoding technology has ensured the recognition speed.

2. We use the flexible hanger-chain segmentation to get flexible software.

As it can be seen . the data after decoding is the coorinate of BRL (not discrete grid). Therefore. the characters can directly be seperated on coordinate data by using hanger-chain algorithms . and their first features can be extracted at the same time.

Generally speaking . each character is a connectivity entity. SO . characters can be seperated by using connectivity. In LAG . we have defined some topological structure relationship and geometrical structure features among BRLs . Base on those defination . we can check the relations among BRLs in LAG while the raster by raster scanning to the BRL is done from top to bottom and from left to right.

. The adjoin BRLs forms a same chain.

. Except the adjoin BRLs . different BRLs form different chains.

. If different chains have relations (nodes or branch dots) . they are defined as a same chain group.

. If different chains have no relations . they form different chain groups.

. Different chain groups form different numeral.

So. after the scan has been done. characters are seperated by checking different chain groups . and the first features can be extracted by recording the first features of adjoin BRL line by line.

The soft segmentaton has no strict requirment for writers and input formats . so the characters can be recognized without marks . and either the scale of characters or the format of handwriting can be no restricted.

3. According to a simple method . two dimension feature can be inflected to one dimension encoding chain. So the lengths of feature is deduced. and the space is thus economized.

4. According to the lengths of the feature the dictionary is devided into eight groups . Take the advantage of the binary-search method to realize the patten matching.

The software is all made in Turbo C language. In the PC/AT computer. therecognition speed is matched to the input speed of fax . that is 4 pages/minute. And the recognition accuracy to the genenal handwritting numbers is 99.9%.

Besides. on some special occasions. all software algorithms can in the ROM of 8086 single-board-computer so that the whole OCR system can be used as a terminal.

Refences

[1] T. Pavlidis : Structural Pattern Recognition.

[2] K. S. Fu :Syntactic Pattern Recognition . Applications.

[3] Heinrich Niemann : Pattern Analysis.

[4] M. Richetin and F. Vernadat : Efficient Regular Grammatical Inference For Pattern Recognition.

[5] B. Duerr. W. Haettich. H. Tropf and G. Winkleb : A Combination of StstisticalAnd Syntactical Pattern Recognition Applied To Classification Of Unconstrained Handwritten Numerals

[6] International CCITT G2/G3 Facsimile Transceiver UF-915 Service Manual.