CHINESE CHARACTER RECOGNITION: A TWENTY-FIVE-YEAR RETROSPECTIVE

G. Nagy
Rensselaer Polytechnic Institute
Troy, New York

ABSTRACT

Twenty-five years ago, in search of both practical solutions and scientific understanding of the problems entailed in computer recognition of a large variety of pattern shapes, the authors undertook the first investigation into machine reading of printed Chinese characters. This early work in the then new area of pattern recognition has since been followed by a variety of research and development projects, mainly in Asia. In this selective survey we attempt to assess current status in the field, and to place the problem of Chinese recognition into perspective compared with other areas of optical character recognition.

INTRODUCTION

The authors were among the first, twenty-five years ago, to conduct large-scale experiments on the classification of machine-printed ideographs [1] . A kindly reviewer wrote that "This well-organized and clearly written paper will be a 'must' reference for the researchers on the automatic recognition of Chinese characters for some time to come." [2] . Since then, hundreds of papers have appeared on this topic, which has assumed both scientific and commercial importance. Although the authors have not directly participated in this research, they have been active in related areas - including the recognition of printed, typed, and handprinted characters, and document format analysis - and therefore hope that their observations today may help further progress in the field.

After a brief review of our early experiments to provide a baseline, we indicate sources of more up-to-date information, including review articles. We then discuss advances in computer technology that have had a significant impact on the problem, and present a sampling of relatively recent research on the classification of both printed and handprinted ideographs. Included in the discussion are techniques of preprocessing (character location and segmentation) and hierarchical classification. On-line character recognition using a graphic tablet or equivalent device, though interesting, is excluded from our survey. We focus primarily on developments outside China, since many of the participants at this conference will be well acquainted with research here. In summary, as a kind of authors' aside that we hope will be thought-provoking and useful to the pattern recognition community as a whole, we attempt to summarize what has been learned about Chinese character recognition as an abstract problem.

In view of space limitations of the conference proceedings, this paper is condensed from a more complete survey which will be published separately.

BACKGROUND

Chinese character recognition c. 1966

The following description of our experiments is quoted from William Stallings' widely cited survey [3].

> "The earliest reported attempt at printed Chinese character recognition is that of Casey and Nagy. The authors used one of the simplest of all pattern recognition techniques: template-matching.
>
> . . . . . .
>
> With the large character set of the Chinese language, template matching becomes very time-consuming. To avoid this, the authors evolved a two-stage matching process. The 'individual' masks, one for each character, are partitioned into groups containing characters with similar masks. Each group is assigned a 'group' mask. In this revised process, a given sample character is compared to the group masks and a preferred order of search through the groups is defined by the mismatch scores. Then, the character sample is compared to individual masks in the established sequence until a sufficiently good match is found. The two-stage approach reduces the average number of comparisons for a sample by a factor of ca. 5-10.
>
> The grouping of characters chosen by the authors is the traditional radical grouping. A radical is a subpattern common to many characters. To obtain a group mask, the matrices of a number of samples of each character in the group (each character with the same radical) are summed and the fifty positions in the sum matrix showing the greatest incidence of black points are used.
>
> Individual masks are designed to discriminate as reliably as possible among characters in the same group. To begin with, each character is assigned a template with thirty defined points selected at random from among those points that are 'stable' for a given character. A black or white point is stable for a character if it consistently appears in a number of samples. Individual masks are then refined by adding template points to improve discrimination.
>
> With their approach, the authors reported an error rate of 1.2%, and a rejection rate of 7%. By varying the mismatch threshold for acceptance, these numbers can be altered, but small improvements in the error rate result in large increases in the rejection rate and vice versa. Improvements in both rates would require higher resolution and hence higher processing time.
>
> The authors are confident that their approach can compete favorably with more sophisticated tech-

niques. This is because of the simplicity of the technique plus the small number of comparisons that have to be made with a two-stage approach. The authors estimate that the proper hardware, including a high-quality scanner, will produce acceptable recognition rates with processing rates of the order of 3400 characters/min."

The work described accurately and succinctly above was motivated only partly by practical concerns, such as input to automatic translation systems. In 1963 digital computing itself was an infant field whose potential was not completely known. Pattern recognition was an application that promised to unleash the full power of "electronic brains". Chinese character recognition offered a challenge that was in certain key aspects representative of the larger world of pattern recognition.

Chinese recognition offers a level of complexity that can not be resolved by direct application of human "common sense". In 1963 OCR design for Latin alphabets was an interactive process. Designers devised meaningful features or prepared templates, then revised them based on recognition tests. Throughout the design period they drew on their knowledge of the character shapes, and of distinguishing differences between similar shapes. This was a practical and effective approach for an alphabet of less than 100 characters.

A Chinese character set of useful size, 1000 symbols minimum and preferably 2000 or more, does not offer this option. The human brain can not hope to assimilate a million or so potential pairwise confusions. There is no recourse but to devise automatic means for determining the distinguishing characteristics of classes.

Our director at IBM Research further saw to it that human preconceptions were not brought to the problem. He assigned a Hungarian-Canadian and an Irish-American to the project. We were forced to consider the Chinese alphabet as a collection of arbitrary shapes, and to develop techniques that could be applied to any large collection of printed symbols. We eventually hedged on this point and sought Chinese counsel. The result was the use of special templates to detect radicals for subgrouping. Nevertheless, we consider the complexity issue fundamental to our efforts, and as we note in our summary, fundamental to the subsequent work of others.

Sources of information on Chinese character recognition

We found about two dozen papers on Chinese OCR in the proceedings of the immediate predecessors of this conference: ICPR 5, 6, 7, and 8. There were a like number of papers in the the proceedings of the 1985, 1986, and 1987 International Conference on Chinese Computing (ICCC). Most of these papers are, however, quite short: more complete descriptions of methods and experiments appear in IEEE-PAMI and IEEE-SMC, in J. PATTERN RECOGNITION, in COMPUTER VISION, GRAPHICS AND IMAGE PROCESSING, and in J. IECE (Electronics and Communications in Japan). We may also expect to see relevant contributions in COMPUTER PROCESSING OF CHINESE AND ORIENTAL LANGUAGES (published since 1983 by the Chinese Language Computer Society) and in COMPUTER SCIENCE AND TECHNOLOGY (published jointly by Science Press, Beijing and Allerton Press, New York).

In addition to the survey mentioned above, Stallings wrote a valuable article on the morphology of Chinese characters, including formal linguistic models and representations [4] . A good bibliography on pre-1973 research was included in a paper presented by Paul P. Wang at the First International Symposium on Computers and Chinese Input/Output Systems (Part I), and at the First International Joint Conference on Pattern Recognition (Part II). A condensed version of his research with C. Shiau was published in PATTERN RECOGNITION [5].

A comprehensive set of references to recent work in China, including a system that can recognize 3000 printed characters at 99% accuracy using a one or two stage preclassifier, can be found in [6]. Kenichi Mori's surveys - Mori was one of the architects of the very successful Japanese postal-code reader - are excellent [7] . Shunji Mori's little book on pattern recognition has a strong emphasis on OCR and he illustrates most of the principal approaches in use [8] .

For those wishing to follow the evolution of the field from the perspective of single researchers, the sequence of papers published by Kazuhiko Yamamoto *bl*9 – 11*br* is illuminating. Finally, Ching Y. Suen's comments on Chinese character recognition in [7] and his brief survey in the Proceedings of the 1983 International Conference on Text Processing with a Large Character Set [12] are well- worth reading: he has probably performed the most extensive experiments in the West.

Technological developments

First and foremost is, of course, the rapid increase in computing power and storage. Experiments equivalent to our early tests can now be conveniently be performed on personal computers (which are far more powerful than the IBM 7094 we used in 1963), while others, for example large-scale clustering of characters, can be carried out on supercomputers.

An important factor is the development of high-resolution matrix printers and displays, without which Chinese I/O is virtually impossible. As in the West, the proliferation of electronic printers, each with its own specially designed font set, increases the range and variability of machine printing to be recognized. Large printed data sets can now be generated in a variety of formats. Nevertheless, handprinted characters still appear to play a more dominant role in the East than in the West, and most of the papers we have seen focus on the far more difficult problems presented by such material.

Consequently, an extremely important development is the availability of very large, segmented and labeled handprinted Kanji databases, such ETL-8 (160 sets of 950 handprinted educational characters) and ETL-9 (200 sets of 3,036 handprinted Japanese Industrial Standard Level-1 characters digitized into 63 x 63 binary pixel arrays, 4000 authors) prepared by the Eletrotechnical Laboratory [11] . Many of the Japanese researchers make extensive use of these samples. The preparation of such test sets is greatly facilitated by the availability of improved Chinese input systems: in our experiments, identification of the sample set proved to be a major problem.

Commercial OCR devices for single-column typewritten or printed pages, using PC software or special processor boards, are now relatively inexpensive. Accordingly, more research has been focussed on relatively complex, multifont material such as technical articles and newspapers in both English and Japanese. Higher-priced OCR machines using dedicated microprocessors are capable of interpreting such pages with relatively little human post-editing.

The lower cost of solid-state scanning arrays has fostered

rapid progress in page digitizers with considerably higher resolution than ordinary television cameras. 240 and even 300 pel/inch digitizers used in OCR of Latin fonts are probably inadequate for small Chinese print, but 400 pel/inch and higher are readily available. Research has also continued on dynamic thresholding methods which allow scanning print with low contrast; indeed, the debate continues on the appropriate trade-off between higher- resolution black/white and lower-resolution grey-scale for both scanning and display. It is widely recognized, however, that 32 x 32 bit arrays do not adequately represent either printed or handprinted Chinese characters: 64 x 64 resolution is now common.

Further codification of Chinese character sets is discussed in [13] (which also has a good review of Chinese and Japanese keyboards). The two important items are the continuing legislatively mandated simplification of the ideographs themselves, and the standardization of the computer codes associated with each ideograph (i.e., the equivalents of ASCII and EBCDIC for Chinese characters). Without coordination, the entry, transfer and display of Chinese text on computers of different manufacture would be impossible. The lack of standard codes would also preclude the wide availability of standard labeled test sets. The number of Chinese characters used in various settings in Japan, and some early experiments on typed Chinese characters, are discussed in [14].

Among significant advances in methodology, we note the development of hierarchical classification schemes far more granular than our old two-level system. Tree classifiers select the pixel or feature most appropriate at every step of the method, so the entire vocabulary is divided at each step of the process.

SELECTIVE SURVEY OF PROGRESS

Preprocessing

We consider briefly three aspects of preprocessing: thresholding to obtain "good" binary representations of the patterns; line and character location including character segmentation; and complex format analysis.

The thresholding of grey-level pictures into binary arrays is studied in [15]. The complexity of the binary image rises, falls, rises again, and falls to zero as the threshold is varied from one extreme to the other. It is conjectured that the optimum is in the valley between the two peaks. An entropy-based measure of complexity is used. Nevertheless, simpler methods, such as the popular Otsu algorithm in [15] are still more commonly used when the scanner does perform the thresholding in hardware.

Various profile-based methods for the segmentation of a line of hand-printed characters are compared in [16]. Each character is coded with the identity of the successful methods. Segmentation using context ("key" characters in Kanji postal codes) is described in [15] . In experiments on 1200 samples, context improves segmentation accuracy from 80% to 90%. According to Murasea, about 10% of handprinted characters are incorrectly segmented without context: he tests an interactive method on 20 sets of 2148 characters.

A page scanned at 200 pels/inch is divided into three kinds of components: text-line, figure/table, and noise in [17] . The method uses an algorithm for connected components and vertical projections. The thin vertical rectangles which may enclose only part of a character are merged.

Akiyama uses three basic features to subdivide a page into meaningful regions: projection profiles, crossing stroke counts, and the enclosing rectangles of connected components. The problem is compounded by the alternation of horizontal and vertical headlines in Japanese newspapers. Profiles are also used in format analysis by M. Tanaka, who estimates that the system, based on known periodicities, is able to detect 99.8% of the text lines.

Features

A review of features used to reduce the dimensionality of high-resolution Chinese characters appears in [18]. Suen also compares experimentally several homogenous sets of features. 500-2000 samples are classified by tree-classifiers constructed from each of the feature sets. The methods considered are (1) Fourier transforms of average grey-levels (cell histograms), (2) elastic profiles, (3) condensed patterns with angular projections, and (4) crossing counts. The synthetically generated distortions are based on a non-white noise model, but the 98.7% recognition rate obtained may not be sustainable on scanned characters.

Another experimental comparison of features (stroke following, "window" method using chain-code tracing, and a modified Hough transform) are compared in [19] , but the comparison is strictly qualitative, because a classifier was not implemented.

Shunji Mori compares a number of techniques for extracting primitive stroke segments (an idea he attributes to Kasvand and Kobayashi) . He advocates a shrinking method based on directional 3x3 spatial filters, which is similar to erosion with a fixed kernel. The extraction is hierarchical, with the most reliable strokes extracted first.

The extraction of radicals by background thinning is demonstrated on a small part of the 881 Chinese character subset of the ETL-8 database [20]. The characters tested has 0,1, or 2 radicals: only left/right division was considered. A success rate of only 85% was obtained, but the resulting errors were carefully analyzed. The segmentation of individual characters into sub- rectangles in this case is 89% accurate.

Hierarchical methods

Using the four-corners method, with a grammar based on horizontal and vertical crossings, it is possible to preclassify 6763 characters into 1586 groups with 97.5% accuracy [21].

The idea of font classification, which has proved useful on Roman characters, is applied to Kai, Soong, and bold fonts in [22]. The experiments are, however, restricted to patterns generated by a synthetic noise generator from only 500 samples.

A complex multilevel classification scheme based on the 4C (corners) and 4P (sides) codes, which can (ideally) encode 4096 classes with 1-6 characters in each class, is discussed in [23]. The experiments are on three groups of 4-6 "complex" characters. After readjustments on several passes on the same characters, 99.9% recognition is obtained, which is then (perhaps prematurely in the opinion of the reader) claimed for the method.

Kimura provides statistics on the number of classes generated and the computational effort versus various similarity measures [24]. The clustering is intended to serve in a preclassification stage.

Experiments on printed characters

A large fixed-pitch single-font (either Min, Black, or Typewriter) experiment with the 5401 classes of the Standard

Interchange Code is reported in [25]. The features, which are extracted by special-purpose hardware, are the number of crossings of rays. The correct candidate has the highest score 95% of the time, and appears among the top four 97.5% of the time. This machine is capable of classifying 300 characters/minute: because its output is a speech synthesizer, an extremely low error rate is not required. The most significant source of errors is mis- segmentation.

A number of major Japanese manufacturers (NTT, Ricoh, Toshiba, Sanyo, Matsushita, Fuli Electronics, and Oki) market systems for printed characters. These systems advertise recognition rates of more than 99% with speeds ranging from 5 to 30 characters per second. The faster ones are stand-alone systems, the slower ones are PC-based: digitization is at 300 or 400 pels per inch.

Experiments on handprinted characters

Partially linear, partially quadratic discriminants in either 16- or 64-dimensional features based on the 4-D codes extracted from the ETL-8 database are compared to a variety of simpler distance measures with good success in [24] . A recognition rate of 98.5% is achieved

using 956 classes in the smaller space and 271 classes in the larger space. The key seems to be the improved estimators for the covariance matrices.

Yamamoto obtains 98.5% correct recognition on one "alphabet" of the ETL-9 set using complex relaxation methods. Using locally minimized correlation and a blurring technique, Saito obtains only 90% on four sets of 952 characters, while Shiono shows the improvements possible with a simple template matching method if the number of templates per character is significantly increased.

SUMMARY

In this summary we offer observations in which we assimilate our own early efforts with later work. In doing this we seek to state general conclusions about Chinese character recognition considered as a model pattern recognition problem.

Knowledge of structure of Chinese characters.

Overall, the articles surveyed present a vast array of techniques for characterizing patterns that experience variations typical of printing or handwriting processes. The characterizations themselves are broad in scope, and can generally be applied to other classes of patterns as well as to Chinese. In particular, research on recognition of handprinted Latin characters can gain insight and useful techniques by studying work on Chinese. This verdict would seem to vindicate the judgment of those who steered us towards Chinese recognition originally as a very general form of pattern recognition problem.

Role of technology

Technology has made Chinese recognition commercially practical, but good feature extraction and classification algorithms are nevertheless the key to a successful system. This is perhaps an obvious point to workers in the field, but still worth recording. Faster processing, large-scale storage, sensitive high-resolution scanners - all orders of magnitude lower in cost than in 1963 - have not in themselves supplied the answers to recognizing Chinese or other alphabets. One still cannot simply store all possible pattern variations as templates. The degree of success depends on the aptness of the sequence of processing algorithms from preprocessing onwards.

Multilevel classification

For efficiency, systems for Chinese OCR are invariably organized as a hierarchy of classifiers. However, it is at first glance puzzling that only two types of hierarchies have been explored. Some workers have designed many-level decision trees, but most research has been directed at a 2-stage approach (as was ours). The effect of the latter is to reduce the candidate set for second-stage classification to an alphabet of fewer than 100 characters, about the same size as a typical Western alphabet. Why do we not find 3- and 4-level schemes? If there were a million classes instead of a few thousand would a two-stage process still be preferred?

We see this dichotomy as the logical outcome of maximizing thruput in two different approaches to classification. The two possible system organizations are: (1) Every feature in the feature set is computed for an input pattern. A subset is then compared against reference data for each character in the alphabet in order to determine a candidate set. Another feature subset is used to reduce the candidate set further, and so on, until at some level the remaining features are used to classify candidates. (2) Candidate groupings or "superclasses" are predefined and stored in a structure that may be viewed as a tree. A subset of the features is used to determine the branch of the tree followed at any level. Eventually the procedure branches to a leaf that contains the actual character class assigned to the input.

The time required for these methods is the sum of the time spent evaluating features on an input pattern, plus the time spent in comparing these feature values against reference data, i.e., in determining the candidate set. In method (1) the feature evaluation time is fixed, since every feature is used. There is thus little to be gained by extending the hierarchy beyond the point where the time for reference comparison is much less than the time for feature extraction. Since two levels is sufficient to reduce the former by a factor of at least ten, it appears that with Chinese characters as inputs there is no practical advantage to extending the hierarchy further.

In method (2) the advantage is in extending the hierarchy as far as possible, since in the limit one evaluates only the features needed to classify a given input, and the number of superclasses to be considered at each level is minimized as well. The sticking point here is that a tree-growing algorithm must be devised to determine the features to be used and to define the candidate classes at each stage of the hierarchy. This calls for sophisticated analytical tools that were not available to us in 1963. Our system actually was of type (2), since candidate classes were predetermined, but we evaluated all features for each input. The net effect was a time per character equivalent to a system of type (1). We merely used the hierarchy to reduce classification time below feature extraction time.

Postprocessing

In machine recognition of Latin fonts it is common practice now to use computerized lexicons to flag potential classification errors. Spelling checkers have even been used as basic elements of the recognition process. On the other hand most symbols in Chinese and Japanese text represent words. Thus, word-by-word validity checking is useful only if the context is known, e.g., in reading addresses or names within an organization. However, researchers on Chinese recognition have actually gone beyond western researchers, employing the more complex methods of syntax analysis in order to detect errors in text.

CONCLUSIONS

A number of uncited papers that we have studied propose ad hoc methods that are not tested at all or else tested inconclusively - for example, on a few hundred classes with a few samples in each class. Lacking a wholly satisfactory model of digitized Chinese characters on which to base mathematical analysis, we cannot evaluate such methods and we must await more extensive experiments to pass judgment.

In making this judgment, we are guided by the rule of thumb that the number of samples necessary to obtain a reasonably dependable estimate of the error rate on new (i.e., not used in the design) samples is ten times the inverse of the error rate. For example, if we wish to proclaim an overall error rate of 0.1%, we would have to test at least 10,000 characters. Another way of putting it is that testing should proceed until ten errors have been made on each class. Since the errors are not evenly distributed, but concentrated on the most similar pairs, this is not a conservative approach. Furthermore, one error per thousand is hardly acceptable for most applications: each page could have several errors.

ACKNOWLEDGMENT

Dr. E. N. Adams of IBM initially entrusted us with the task of classifying Chinese ideographs. Assistance of the following people in providing input to this survey is gratefully acknowledged: H. Takahashi of IBM Tokyo Research Laboratory, X. Kanai of RPI.

REFERENCES

[1] R. G. Casey and G. Nagy, Automatic Recognition of Machine Printed Chinese Characters, IEEE-TEC, 1966.

[2] R. Bledsoe, Review of Casey and Nagy, "Automatic Recognition of Machine Printed Characters," Computing Reviews, 1966.

[3] W. Stallings, Approaches to Chinese Character Recognition, Pattern Recognition 8, pp 87-98, 1976.

[4] W. Stallings, The Morphology of Chinese Characters: A Survey of Models and Applications, Computers and the humanities (Pergamon) 9, pp 13-24, 1975.

[5] P.P. Wang, The topological analysis, classification, and encoding of Chinese characters for digital computer interfacing - Part I, Proc. 1973 Int'l Symp. on Computers and Chinese Input/Output Systems, Aug. 1973, Academia Sinica, Taiwan, China.

[6] L.-D. Wu, J.-W. Tai, Some advances of pattern recognition in China, ICPR8, pp. 134-143.

[7] C.Y. Suen, Character recognition by computer and applications, Handbook of pattern recognition and image processing, pp. 569-586, Academic Press.

[8] Shunji Mori, Fundations of character and pattern recognition, (book, in Japanese).

[9] Kazuhiko Yamamoto, Recognition of Handprinted Characters by Convex and Concave Features, ICPR5, 708-711.

[10] K. Yamamoto, H. Yamada, T. Saito, R.-I. Oka, Recognition of handprinted Chinese characters and Japanese Cursive Sillabary, ICPR7, pp 385-388.

[11] K. Yamamoto, H. Yamada, T. Saito, I. Sakaga, Recognition of handprinted characters in the first level of JIS Chinese characters, ICPR8, pp 570-572.

[12] C.Y. Suen, Computer recognition of Kanji characters, Proceedings of 1983 Conference on Text Processing with a Large Character Set, pp. 429-435 Chinese Language Computer Society.

[13] Helena Wong Gin, On the Future of Chinese Data Processing: A User Perspective, ICCC6, 365-368.

[14] I. Nakano, K. Nakata, Y. Uchikura, A. Nakajita, Improvement of Chinese Character Recognition Using Projection Profiles, First Int Joint Conf On Pattern Recognition Washington, 1973.

[15] N. Babaguchi, M. Tsukamoto, T. Aibara, Knowledge and character segmentation from handwritten document image, ICPR8, pp 573-575.

[16] Li Bin, Zhao Shuxiang, An Application of Image Segmentation to a Kind of Pattern Recognition Problems, ICCC7, 142-145.

[17] K.-I. Maeda, Y. Kurosawa, H. Asada, K. Sakai, and S. Watanabe, Handprinted Kanji Recognition by Pattern Matching Method, ICPR6, 789-792.

[18] C.Y. Suen, Y.Y. Tang, Q.R. Wang, Feature Extraction in the Recognition of Chinese Characters Printed in Different Fonts, ICCC6, 136-143.

[19] Fang-Hsuan Cheng, Wen-Hsing Hsu, Radical Extraction by Background Thinning Method for Handwritten Chinese Characters, ICCC7, 175-182.

[20] Fang-Hsuan Cheng, Wen-Hsing Hsu, Three Stroke Extraction Methods for Recognition of Handwritten Chinese Characters, ICCC6, 191-195.

[21] G. Chunbiao, X. Bourong, Automatic recognition of printed Chinese characters by four corners method, ICPR8, pp. 1013-1015.

[22] Q.R. Wang, C.Y. Suen Y.Y. Tang, Application of a Statistical Equivalent Block Classifier in the Recognition of Chinese Characters Printed in Different Fonts, ICCC5, H-2.1-H-2.13.

[23] Jun S. Huang, Ma-Lung Chung, Separating Similar Complex Chinese Characters by Walsh Transform, ICCC7, 187-191.

[24] F. Kimura, T. Harada, S. Tsuruoka, Y. Myjake, Modified quadratic discriminant functions and the application to Chinese character recognition, ICPR7, pp 377-380.

[25] High-Way Kuo, Song-Kune Su, Tien-Cheu Kao, Implementation of a Chinese Reading System, ICCC7, 200-203.