Optical Character Recognition: An Overview and an Insight

Deepa Berchmans
Dept. of ECE,
PRSCET
Thiruvananthapuram, India
deepaberch@gmail.com

S S Kumar
Department of EIE
Noorul Islam University
Thuckalay, Nagercoil, Tamil Nadu, India
kumar_s_s@hotmail.com

Abstract—**Many researches are going on in the field of optical character recognition (OCR) for the last few decades and a lot of articles have been published. Also a large number of OCR is available commercially. In this literature a review of the OCR history and the various techniques used for OCR development in the chronological order is being done.**

Keywords— Optical Character Recognition, intelligent character recognition, handwritten characters, printed characters, document analysis, glyph recognition, pattern recognition, online recognition, offline recognition.

I. INTRODUCTION

After the advent of digital computers, incorporating human functions to computers has been an interesting and exciting research field [1]. For over years, humans have been thinking of machines with the ability to "read" and interpret printed textual documents,so that they can be automatically converted into an alternate medium or format [2]. Efficient algorithms have been developed so far so as to enable the machines to recognize characters. Such a system is named as Optical Character Recognition (OCR). This is a system developed for deriving character-based files from digitized images of printed or typewritten documents and/or handwritten manuscripts. Digitizing is done by using flat bed scanners or digital cameras. It is thus a process of visual recognition, which converts text documents into editable/searchable text.

OCR technology improves the efficiency of office work to a greater extent as, when done accurately the transformation enables searching the electronic copies of documents that might otherwise remain tucked away in filing cabinets gathering dusts. Because of this advantage of the OCR systems, they are employed in a wide area from banking to archaeology.

Based on the input device, there are two classes of character recognition: on-line and off-line recognition. The first category systems uses devices like digitizer tablets for data acquisition and in this system, recognition is done while writing it. On the other hand, the latter systems collect data from static devices such as scanners and cameras. On-line recognition system is real-time due to the concurrent data collection structure, whereas off-line recognition system requires specific methodologies for preparing the image prior to recognition, which includes noise removal and restoration of the input image to eliminate the damage caused during the acquisition process.

Recognition of machine-printed and clear handwriting characters (HCR- Handwritten Character Recognition) has almost been achieved and there are commercial systems in the market. Even then recognizing unconstrained handwritten characters and determining the similarities between some characters and that too in poor quality of paper are the still challenging problems to be resolved. Thus, the issue of similarity between the characters, broken and distorted characters and a number of varieties of character shapes is to be addressed by HCR systems. Thus, substantial researches are being carried out to develop a system with intelligence to recognize the natural handwriting with minimal errors in almost all the scripts around the world.

II. HISTORY OF OCR

The idea of character recognition was first introduced by the invention of retina scanner, an image transmission system which makes use of mosaic of photocells [3]. A major breakthrough for the today's television and reading machines happened in 1890 with the invention of a sequential scanner by Nipkow. During the early ages OCR was considered as an aid for the blind people. But later on it developed into a vast field of research and development.

The first evidence of optical character recognition system is a patent filing of Tauschek in Germany in 1929 [7] and later he was given US Patent in 1935 and Handel was given independently in 1933 [8]. Both the machines use a circular disc with template symbols cut out of it so that light shines through it. The image to be recognized is held in front of the disc and is then illuminated. The light reflecting off a portion of it is then focused through the template hole and detected by a photosensor.

The commercially available OCRs may be classified into four generations based on their efficiency, robustness and versatility. In the first generation OCRs could read only selected fonts and shapes of characters. Such machines were employed in early 1960s. The first generation OCR to become widely commercialized was IBM 1418 [4]. The technique used was logical template matching. The second generations OCRs were much more capable, which could recognize both

978-1-4799-4190-2/14/$31.00 ©2014 IEEE

characters printed by machine as well as written by hand. The second generation OCRs which were available during the middle of 1960 to early 1970s was restricted to numerals only. The first OCR system of second generation was IBM 1287, which was a hybrid system that combines both digital and analog technology [4].

Further research were done in the quality of print being scanned and thus developed the third generation OCRs. They operate on hand-written characters of large set than before and poor print quality characters. Such systems were popular during the period 1975-85 ([4]-[6]).

The fourth generation OCRs are capable of scanning and recognizing characters from complex documents intermixed with texts, tables and mathematical symbols and also low quality noisy documents such as fax and photocopies, unconstrained handwritten characters and colour documents.

Currently more sophisticated OCRs are available for Japanese, Chinese Arabic and Roman primer ([3], [9]-[14]).

III. RELATED WORK

Hamanaka et al (1993) [15], suggested a methods that is effective in the recognition of Japanese characters. The conventional methods used till dates restricted the order and number of strokes. The offline methodology removes the above said constraints based on the pattern matching of orientation of feature patterns. It can be improved with the enhancement in nonlinear pattern matching, nonlinear shape normalization, and the normalization-cooperated feature extraction method. The recognition rate attained was 95.1%.

Ohhira et al. (1995) [16], proposed a system using plural combination of Neural Networks and which could automatically recognize 6709 Chinese characters. The system consists of four parts: - rough classification part, fine classification part, recognition part, and auto judgement part. The system operates by classifying the input data by classifying by character density at the rough and fine classification parts. The multi-layered NN recognizes at the recognition part. The auto judgement part judges and output the values. The authors claim 100% recognition efficiency.

Alherbish et al. (1997) [17], introduces a parallel recognition system for Arabic characters. The objective of the system was to simultaneously attain high speed and full precision. The system uses distributed computing and parallel processing techniques to accomplish the goal. This multi-processing system enhances Arabic character recognition systems of that time.

Tanaka et al. (1999) [18] proposes a system for hand written character recognition. The systems integrate offline recognition and online recognition. Offline character recognition requires a bitmap image where as online character recognition requires an input pattern as an arrangement of x-y coordinates. When integrated together, these methods complement each other, as every method has distinct recognition capabilities. The system may be used for both offline and online recognition. A recognition rate of around 73% for offline and 85% for online recognition was claimed.

Ikeda et al. (1999) [19] suggested a technique which uses Hidden Markov Model so as to upgrade the recognition rate of offline character recognition systems. The system proposes an alternative for the complex 2D HMM structure. The difficulty in the structure was that it was very difficult to get samples which could guarantee successful generalization. To get around the issue, a method is put forward for glyph recognition using 1D HMMs in various directions through 2-dimensional feature extraction. With the aid of this approach, the recognition rate is raised by about 1% when compared to the 1D HMM character recognition system.

Bensal et al. (2000) [20] proposed a system that uses a word dictionary tailored for OCR. These are formed by integrating many knowledge sources in hierarchical manner for Devanagari Script recognition. The glyph classification was based on a hybrid technique. Using this word dictionary a classification process is carried out. Based on the results of this classification procedure, a decision will be made whether to perform further segmentation of the image box is to be done or not.

Arica et al. (2002) [21] proposed a method for the off-line cursive handwriting recognition problem. The technique uses a sequence of image segmentation and recognition algorithms. Initially global parameters, such as base lines, stroke width and height, slant angles are evaluated. Then, a segmentation process which segments characters is used. Third, a shape recognition technique to label and rank the characters that is based on hidden Markov model (HMM) is made use of. Finally, to optimize the problem for word-level recognition, information from HMM ranks and lexicon are combined. This method corrects utmost errors produced by the HMM ranking stages and segmentation by maximizing an information measure.

Kang et al. (2004) [22], proposed a system in which the strokes of characters and relationship between characters are represented stochastically. A character/glyph is characterized by a multivariate RV (random variable) over the components and its probability distribution is studied from a training data set. The character is resolved into factors and is almost corrected by a set of lower-order probability distributions. As per the method put forward by the authors, a handwritten Hangul character recognition system was developed which gives better results.

Liu et al. (2005) [23], used several methods of handwritten character recognition on the baseline system. In the proposed system a gradient feature is extracted in the feature extraction stage. This provides high resolution on both angle of the strokes and magnitudes in the glyph image. The efficiency of Modified Quadratic Discriminant Function classifier is finally enhanced in the classification stage by several demarcation schemes, adding minimum classification error (MCE) training on the classifier parameters and modified distance to represent parameters and resembling characters discrimination. Each of these techniques used leads to the enhancement of the rate of character recognition.

Liu et al. (2006) [24], proposed a methodology for the recognition of characters with low resolution. This technique suits the input character for the apt database according to the quality of the input image. It consists of two stems: glyph image quality estimation and glyph recognition. Firstly, it considers the gray distribution feature to evaluate the glyph image quality. Then, as per the estimation result, the suitable glyph database and the recognition method are selected for the input image which makes the classification have the highest likelihood of being the precise decision.

Huang et al. (2007) [25], put forward a method of radical-based online recognition of Chinese handwritten characters using support vector machine (SVM). The input characters are pre-processed segmented and feature extracted. Then in order to arbitrate the type of the pattern of the glyph, the midpoint of each segment is projected in vertical and horizontal directions. The glyph is thus disintegrated into significant sub-structures. Every substructure is a radical and is split into 8 subareas for all the four directions so that the statistics feature of number of pixels in each subarea is suited to recognize radical using SVM. The recognition of Chinese character is converted to a series of radical matching between the input glyph and reference pattern. The front or rear radical is utilized in coarse classification stage to reduce the number of candidate glyphs. The structure and statistical features of characters are also taken on in this technique. This method is independent of stroke number and stroke order of characters.

Kannan et al. (2008) [26], proposed a system for offline recognition of Tamil characters written in cursive handwriting. The system utilizes a combination of frequency domain and time domain feature of characters and is based on HMM. The authors claim the system to be flexible at the same time robust. Higher degree of accuracy in results has been obtained by carrying out thus technique on a global database.

Prasad et al. (2009) [27], put forward a technique to recognize characters from Gujarati language. They suggest a method known as pattern matching in which the glyph is distinguished by analyzing its shape and comparing its features that discriminate each and every character. Prior to recognition pre-processing and image enhancement is done. Several characters from handwritten forms or peripheral devices are recognized using neural networks.

Sobu et al. (2010) [28], put forward a binary tree-based clustering process which is capable of keeping the accuracy as high as possible. It has been experimentally proved that the character recognition using this clustering technique is 8.3 times quicker than the full linear matching at mere 0.22% precision drop. This rate can be enhanced to 36.2 times, by the combined application of Sequential Similarity Detection Algorithm (SSDA) and a PCA-based dimensionality reduction.

Kumar et al. (2011) [29], presented a scheme for offline recognition of characters written in Gurumukhi script. This method is based on k-NN classifier. The feature information concerned with the character is extracted initially and a skeleton of the character is prepared. This is based on the transition and diagonal features of the Gurumukhi character. Based on the allocation of points on the bitmap image of the character, the transition and diagonal features are computed. The Euclidean distance from testing point to the reference points are measured so as to to locate the k-nearest neighbours.

Yutao et al., (2012) [30], proposed a hybrid system based on Generalized Regression Neural Network which employs wavelet transform. To extract the features from input Chinese characters, a wavelet transform based block projection adopted. In order to minimize the dimension of feature vector, a cluster algorithm is introduced further and to identify similar characters an approach of regional recognition is also introduced in this system. A GRNN with significant non-linear mapping ability and improved fault-tolerant capability is developed as character classifier. The classification algorithm has more robustness than the algorithms introduced so far.

Patel et al. (2013) [31], presents a method of handwriting character recognition. To enhance the accuracy of recognition at the pixel level, computational capability of Euclidean distance metric and the learning capability of artificial neural network, this method utilizes the compression capability of discrete wavelet transform. The problem of handwritten character recognition has been addressed with multi-resolution technique using discrete wavelet transform and learning rule through the artificial neural network. Handwritten characters are categorized into 26 pattern classes based on apt properties. During pre-processing each character is captured within a rectangular box and then resized to a threshold size. The learning rule of artificial neural network is having been used for computing the weight matrix of all classes and then recognition scores are generated by fusing the unknown input pattern vector with the weight matrices of all the classes. Utmost value of the score corresponds to the identified input character. Greater recognition rate was achieved with this method.

Prum et al. (2013) [32], introduced a technique for character recognition based on Support vector machine, which used a discriminative method based on explicit grapheme segmentation. Both single character recognition as well as bi-character recognition is possible with this method. Bi-character refers to shared character parts, whose recognition was practically difficult till that time. Whole word recognition is achieved with an efficient dynamic programming method.

Giri et al. (2013) [33], utilizes an algorithm which is very close to human recognition system of objects. This technique recognizes characters by analyzing minimal amount of data. The method is independent of the size of the character being recognized.

978-1-4799-4190-2/14/$31.00 ©2014 IEEE

Dassanyake et al. (2013) [34], proposed Handwritten Character Recognition system which is implemented with the capability of extracting the content of an image. A background process is run by the conversion process, without any involvement of the user. User can perform the editing of the converted text after completing the conversion, in the Panhinda editor. This document describes the techniques for enhancing the quality of the image, character segmentation, character recognition and digital dictionaries. Noise removal, lighting conditions and angle effects are done at the pre-processing phase. Character segmentation is done after obtaining a binarized image using Horizontal and Projection Profile method. For recognizing the characters, the Support Vector Machine technique was used. Error correction is done by using a combined model of noisy channel model and natural language model.

IV. CONCLUSION

In this paper, a survey of the progress being done in the area of OCR over the past twenty years has been discussed. The paper is organized in the chronological order of year and discusses the works starting from the year 1993. The different techniques developed for the implementation of handwritten OCR in the International scenario for the past two decades gives an idea of the various techniques that being used and how to enhance and develop them in the future.

V. SCOPE OF FUTURE WORK

Future enhancements may be done for developing OCRs for both printed and handwritten documents in poor quality papers. Also OCRs may be developed for regional languages; even though literatures are available for some languages such as Hindi, Kannada etc. Also research may be done for the development of OCRs for recognizing multi-font or multi-script characters.

REFERENCES

[1] Fischer S (2000). *Digital Image Processing: Skewing and Thresholding*, Master of Science Thesis, University of New South Wales, Sydney, Australia..

[2] George Nagy. "*Twenty years of document image analysis is PAMI*". IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(1): 38-62, January 2000

[3] .J. Mantas, "*An overview of character recognition methodologies*", Pattern Recognition 19 (1986) 425-430

[4] S. Mori, C.Y. Suen, K. Yamamoto, "*Historical review of OCR research and development*", Proc. IEEE 80 (1992) 1029-1058.

[5] S. Mori, K. Yamamoto, M. Yasuda, "*Research on machine recognition of hand-printed characters*", IEEE Trans. Pattern Anal.Mach.Intell. 6 (1984) 386-405.

[6] G. Nagy, "*At the frontiers of OCR*". Proc. IEEE 80 (7) (1992) 1093-1100.

[7] Gustav Tauschek. Reading machine. U.S. Patent 2026329, http://www.google.com/patents?vid=USPAT2026329, December 1935*FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.

[8] Paul W. Handel. Statistical Machine. U.S. Patent 1915993, http://www.google.com/patents?vid=USPAT1915993, June 1993.

[9] A. Amin, "Off-line Arabic character recognition: the state of the art", Pattern Recognition 31 (1998) 517-530.

[10] F. El-Khaly, M. A. Sid-Ahmed, "*Machine recognition of optically captured machine printed Arabic text*", Pattern Recognition 23 (1990) 1207-1214.

[11] T. S. El-Sheikh, R. M. Guindi, "*Computer recognition of Arabic cursive scripts*", Pattern Recognition 21 (1988) 293-302.

[12] L. O' Gorman, R. Kasturi, "*Document Image Analysis*", IEEE Computer Society Press. Los Alamitos, C.A. 1995.

[13] A.J. Rocha, T. Pavlidis, "*Character recognition without segmentation*", IEEE Trans. Pattern Anal. Mach. Intell. 17 (1995) 903-909.

[14] W. Stallings, "*Approaches to Chinese character recognition*". Pattern Recognition Pergamon Press. Vol. 8, pp. 87-98. (1976) 87-98.

[15] Hamanaka, M., Yamada, K. ; Tsukumo, J., "*On-line Japanese character recognition experiments by an off-line method based on normalization-cooperated feature extraction*", Proceedings of the Second International Conference on Document Analysis and Recognition, 1993., 204-207.

[16] Ohhira T., Pecharanin N., Taguchi A., Iijima, N. ; Akima, Y. ; Sone, M., "*Chinese character recognition by the auto recognition system*", IEEE International Conference on Neural Networks, Volume: 5, (1995) 2222-2225.

[17] Alherbish J., Ammar R A, Abdalla M., "*Arabic character recognition in a multi-processing environment*", Proceedings of the second IEEE Symposium on Computers and Communications (1997), 286-291.

[18] Tanaka H, Nakajima K, Ishiqaki K, Akiyama K, Nakagawa M, "*Hybrid pen-input character recognition system based on integration of online-offline recognition*", Proceedings of the fifth International Conference on Document Analysis and Recognition (ICDAR-1999) 209-212."

[19] Ikeda H, Ogawa M, Nishimura H, Sako H, Fujisawa H, "*A recognition method for touching Japanese Handwritten characters*", Proceedings of the fifth International Conference on Document Analysis and Recognition (ICDAR-1999) 641-644.

[20] Veena Bensal, R.M.K. Sinha, "*Integrating Knowledge sources in Devanagari Text recognition system*", IEEE Transactions on Systems, Man. And Cybernetics-Part A: Systems and Humans, Vol 30, No.4, July 2000.

[21] Arica N, Yarman-Vural F T., "*Optical character recognition of cursive handwriting*", IEEE Transactions on Pattern Anal. And Mach. Intell. Volume 24, Issue 6, (2002) 801-813.

[22] Kyung-Won Kang, Kim J H., Utilization of Hierarchical, "Stochastic *relationship modeling for Hangul character recognition*", IEEE Transactions on on Pattern Anal. And Mach. Intell. Volume 26, Issue 9, (2004) 1185-1196.

[23] Hailong Liu, Xiaoging Ding, "*Handwritten character recognition using gradient feature and quadratic classifier with multiple discrimination schemes*", Proceedings of the Eighth International Conference on Document Analysis and Recognition (2005), (vol. 1) 19-23.

[24] Chunmei Liu, Chunheng Wang, Ruwei Dai, "*Low resolution character recognition by Image quality evaluation*", Proceedings of the Eighteenth International Conference on Pattern Recognition (ICPR 2006) 864-867.

[25] Xingiao Lv, Dongshan Huang, ENming Song, Ping Li, CHunshan Wu, "*One Radical-Based on-line character recognition (OLCCR) system using support vector machine for recognition of Radicals*", 1st International Conference on Bioinformatics and Biomedical Engineering, 2007, 558-561.

[26] Kannan R J, Prabhakar R, Suresh RM, "Off-*line cursive handwritten Tamil character recognition*", International Conference on Security technology (2008) 159-164.

[27] Prasad J R, Kulkarni U V, Prasad R S, "Offline *handwritten character recognition of Gujarati script using pattern matching*", 3Rd International Conference on Anti-counterfeiting Security and Identification in communication, 2009, 611-614.

[28] Sobu Y, Goto H, Aso H, "*Binary tree-based precision-keeping clustering for very fast Japanese character recognition*", 25th International Conference on Image and Vision computing, New Zealand, 2010, 1-6.

[29] Kumar M, Jindal M K, Sharma R K, "*k-nearest neighbor based offline handwritten Gurumukhi character recognition*", International Conference on Image Information processing (2011) 1-4.

[30] Wang Yutao, Qin Tingting, Tian Ruixia, Yang Gang, "*Recognition of license plate character based on wavelet transform and generalized regression neural network*", Control and Decision Conference (CCDC), 2012 24th Chinese, 1881-1885.

[31] Patel D K, Som T, Singh M K, "Multi-resolution technique to handwritten English character recognition using learning rule and Euclidean distance metric", International Conference on Signal Processing And Communication, 2013, 207-212.

[32] Prum S, Visani M, Fischer A, Ogier J M, "A *discriminative approach to on-line handwriting recognition using Bi-character models*", 12th International Conference on Document Analysis and Recognition, 2013, 364-368.

[33] Giri K J, Bashir R, "*Design and Implementation of a novel cognitive character recognition technique*", International Conference on Signal Processing and Communication, 2013, 225-229.

[34] Dassanyake D M D S S, Yasara R A D D, Fonseka H S R, HeshanSandeepa E A, Seneviratne L, "*Panhinda-offline character recognition system for handwritten articles*", International conference on IT convergence and security, 2013, 1-4.

978-1-4799-4190-2/14/$31.00 ©2014 IEEE