Review on OCR for Handwritten Indian Scripts Character Recognition

Munish Kumar[1], M.K. Jindal[2], and R.K. Sharma[3]

[1]Assistant Professor, Computer Science Department,
GGS College for Women, Chandigarh, India
[2]Associate Professor, Department of Computer Science & Applications,
Panjab University Regional Centre, Muktsar, India
[3]Professor, School of Mathematics & Computer Applications,
Thapar University, Patiala, India
munishcse@gmail.com, manishphd@rediffmail.com,
rksharma@thapar.edu

Abstract. Natural language processing and pattern recognition have been successfully applied to Optical Character Recognition (OCR). Character recognition is an important area in pattern recognition. Character recognition can be printed or handwritten. Handwritten character recognition can be offline or online. Many researchers have been done work on handwritten character recognition from the last few years. As compared to non-Indian scripts, the research on OCR of handwritten Indian scripts has not achieved that perfection. There are large numbers of systems available for handwritten character recognition for non-Indian scripts. But there is no complete OCR system is available for recognition of handwritten text in any Indian script, in general. Few attempts have been carried out on the recognition of Devanagari, Bangla, Tamil, Oriya and Gurmukhi handwritten scripts. In this paper, we presented a survey on OCR of these most popular Indian scripts.

Keywords: OCR, Handwritten character recognition, online, offline, Indian scripts.

1 Introduction

Nowadays, world is being influenced a lot by computers and almost all the important processing is being done electronically. As such, it becomes important that transfer of data between human beings and computers is simple and fast. Character recognition is a research problem that has been ongoing since the sixties. It stills an active area of research because the problem is complex in nature. Optical Character Recognition (OCR) is the most essential part of document analysis system. OCR is the field of pattern recognition, image and natural language processing. OCR is the recognition of printed or handwritten text by a computer. This recognition gives a significant benefit in order to bridge the gap between man and machine communication. The document analysis and recognition has played and currently playing a major role in pattern recognition research. In general, research on optical character recognition for Indian scripts is ongoing. But till now no solution has been offered that solves the problem

D. Nagamalai, E. Renault, and M. Dhanushkodi (Eds.): DPPR 2011, CCIS 205, pp. 268–276, 2011.

correctly and efficiently. The process of character recognition can be divided into two parts, namely, printed and handwritten character recognition. The printed documents can further be divided into two parts: good quality printed documents and degraded printed documents. Handwritten character recognition has been divided into offline and online character recognition, as shown in figure 1.

Offline documents are scanned images of prewritten text, generally on a sheet of paper. In online handwriting recognition, data are captured during the writing process with the help of a special pen and an electronic surface. Recognition of offline handwritten documents has been an active research area in the field of pattern recognition. Over the last few years, the numbers of laboratories all over the world are involved in research on handwriting recognition.

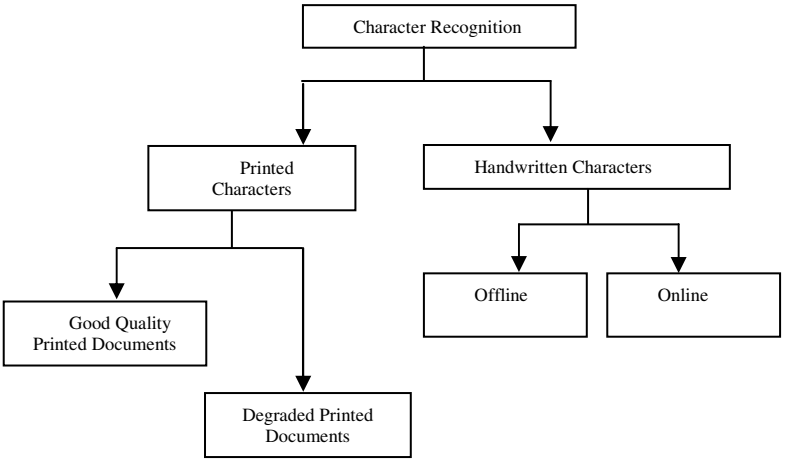

Fig. 1. Different character recognition systems

The recognition of cursive handwriting is very difficult due to large number of variations in shapes and overlapping of characters. In the offline handwriting recognition system, the pre-written document is converted into a digital image through optical scanner. In the handwritten script, there are variations in writing style, size *etc*. Handwritten Character Recognition (HCR) system will enable the computer to process the handwritten documents, which are currently processed manually. One can find these handwritten documents at various places such as post offices, banks, insurance offices and colleges *etc.* for processing data.

1.1 Stages of HCR

A complete process is followed for handwritten character recognition which is shown in figure 2.

1.1.1 Digitization

Digitization is the process whereby a document is scanned and an electronic representation of the original, in the form of a bitmap image, is produced. Digitization produces the digital image, which is fed to the pre-processing phase.

1.1.2 Preprocessing

Preprocessing is used for skew detection/correction, skeletonization, and noise reduction/removal. Skewness refers to the tilt in the bit mapped image of the scanned paper for OCR. It is usually caused if the paper is not fed straight into the scanner. Skeletonization is used for decreasing the line width of text from many pixels to single pixel. Noise removal is used to remove unwanted bit pattern which does not play any significant role in document**.**

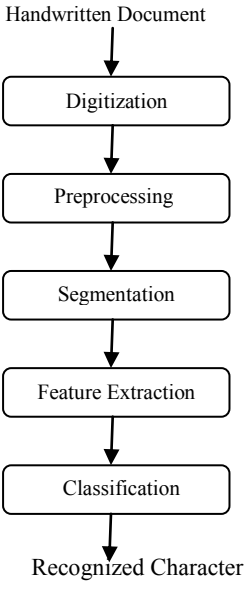

Fig. 2. Handwritten character recognition system.

1.1.3 Segmentation

In the character recognition, segmentation is very important for recognition. Segmentation is used to break the script into lines, words and characters. The challenge of a segmentation technique lies in the decision of best segmentation point for line, word and character isolation. In handwritten script, lots of features are available for segmentation provided by skeleton and the intersection point. Incorrect segmentation can lead to the incorrect recognition. Segmentation of handwritten text is a difficult task owing to variety of writing styles.

1.1.4 Feature Extraction and Classification

Feature extraction is the phase which is used to measure the relevant shape contained in the character. In the feature extraction phase, one can extract the features according to levels of text, e.g., character level, word level, line level and paragraph level. The classification phase is the decision making phase of an OCR engine, which uses the features extracted in the previous stage for making the class memberships in pattern recognition system. The preliminary aim of classification phase of OCR is to develop the constraint for reducing the misclassification relevant to feature extractions.

2 Properties of Indian Scripts

There are 23 languages in India [1] namely Assamese, Bengali, Bodo, Dogri, English, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Santhali, Sindhi, Tamil, Telugu and Urdu. There are 14 different scripts in India Assamese, Bangla, Devanagri, Gujarati, Gurmukhi, Kannada, Kashmiri, Malayalam, Oriya, Roman, Tamil, Telugu and Urdu used for writing these languages. Indian scripts are different from non-Indian scripts in several ways. Indian scripts are composition of symbols like: consonants and modifiers. In Indian scripts case sensitivity is absent. The Indian scripts are divided into three zones as shown in figure 3. As compared to non-Indian scripts, the research on OCR of handwritten Indian scripts has not achieved that perfection. Few attempts have been carried out on the recognition of Devanagari, Bangla, Tamil, Oriya, Telugu and Gurmukhi handwritten scripts [2-5]. There is no complete OCR system is available for commercial use for recognition of handwritten text in any Indian script, in general.

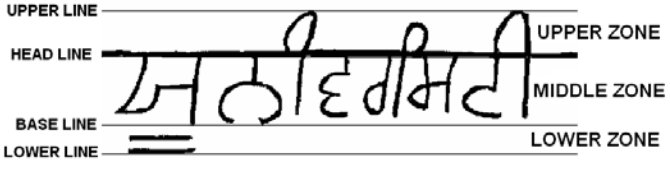

Fig. 3. Different zones of Gurmukhi text

3 Handwritten Character Recognition

The research for developing an OCR system started in the nineteenth century and a system was available in 1929. Modern version of OCR was developed in 1951 by David Shepard. This became very popular owing to its commercial use. Handwriting is the natural mode of collection and storing information for the human beings. It can also be used for communication between human beings and computers. Handwritten character recognition system can be evaluated from several perspectives. One major distinction can be made from the recognition process either during the writing (on-line) or from earlier handwritten document (offline). The recognition of handwritten character is very complex due to non-uniformity in size and style. In general, the location of the characters is not predictable, nor is the spacing between them. An online handwriting OCR was also available during 1950s in which an electronic tablet was used to capture the x-y values of pen movement. Mori *et al.* [6] have divided the OCR into three generations: first generation is for printed documents, second generation for handwritten documents and third generation is for degraded printed text and documents with text, graphics and mathematical symbols.

3.1 Recognition of Devanagari Script

Devanagri, an alphabetic script, is used by a number of Indian languages, including Sanskrit, Hindi and Marathi. Kumar [7] has proposed an AI based technique for machine recognition of Devanagari handwritten script. He has used three levels of

abstractions to describe this technique. Recognition of Devanagri handwritten script was provided in 2007 by Hanmandlu *et al.* [8]. They have proposed the system for recognition of handwritten Hindi characters based upon the membership function of fuzzy sets. Sethi and Chatterjee [9] have also done some work on Devanagari script. They have presented a Devanagari hand-printed numeral recognition system based on binary decision tree classifier. Bansal and Sinha [5] have proposed the technique for complete Devanagri script recognition. In this research, they recognize the character into two steps, in the first step, they recognize the unknown stroke and in second step the character based on these strokes is recognized. Bajaj and Chaudhury [10] have proposed a system for hand-written numeral recognition of Devnagari characters.

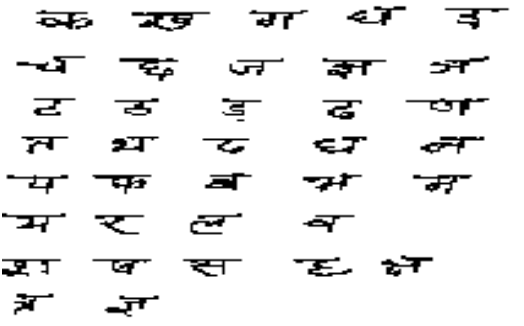

Fig. 4. Handwritten Devanagri characters

3.2 Recognition of Bangla Script

The maximum work for recognition of handwritten characters has been done on Bangla characters. In 1982, Chaudhury *et al.* [11] have proposed a recognition scheme using a syntactic method for connected Bangla handwritten numerals. In this system, the skeleton i.e. structure of character is matched. Pal *et al.* [12] have proposed the technique for Bangla handwritten pin code recognition system. They input the bitmap document and then water reservoir concept is applied to segment and recognize the pin code.

Fig. 5. Handwritten Bangla characters

Bishnu and Chaudhuri [13] have proposed technique for segmentation of Bangla handwritten text into characters based on recursive shape. Dutta and Chaudhury [14] have developed an isolated optical character recognition system for Bangla alphabets and numerals using curvature features. Pal *et al.* [15] have proposed the technique for Bangla handwritten numerals document and then water reservoir concept is applied to segment The recognition system for online handwritten Bangla characters was available in 2007 by Bhattacharya *et al.* [16]. They have used the direction code based features for recognition and achieved 93.90% accuracy from training sets.

3.3 Recognition of Oriya Script

The Oriya OCR system was developed at Indian Statistical institute, Kolkatta are similar to the Bangla OCR system by the Pal and Chaudhuri [17]. They have used the Hough transform based technique for skew angle estimating and recognizing the Oriya alphabets. In 2007, Pal *et al.* [3] have proposed the recognition system for offline Oriya handwritten script. They have used the curvature feature for this proposal and got the accuracy of about 94.60% from few offline handwritten Oriya samples.

3.4 Recognition of Kannada Script

Kannada is one of the major and earliest script of Southern India and spoken by about more than 50 million people in the Indian state of Andra Pradesh, Karnataka, Maharasthra and Tamil Nadu. Little work has been done in Kannada handwritten script recognition. The handwritten Kannada numeral recognition is reported by Acharya et al. [25]. They have used the structural features and multilevel classifiers for recognition.

3.5 Recognition of Malayalam Script

Malayalam script is also one of the script of Southern India and it is the eighth most popular script in India and spoken by about 30 million people in the Indian state of Kerala. The writing nature of Malayalam is similar to Tamil. Rajashekararadhya and Ranjan [26] have proposed the algorithm for feature extraction of Malayalam script recognition. This algorithm can be used for other Southern Indian scripts like Kannada, Telegu and Tamil also.

3.6 Recognition of Tamil Script

Sundaram and Ramakrishnan [18] have proposed the two dimensional principal component analysis (2DPCA) technique for recognition of online Tamil character recognition. The on-line Tamil character recognition is reported by Aparna *et al.* [19]. They have used shape based features including dot, line terminal, bumps and cusp.

3.7 Recognition of Gurmukhi Script

Gurmukhi script is the script used for writing Punjabi language and is derived from the old Punjabi term “Guramukhi”, which means “from the mouth of the Guru”. Gurmukhi script has three vowel bearers, thirty two consonants, six additional

consonants, nine vowel modifiers, three auxiliary signs and three half characters. Gurmukhi script is 14th most widely used script in the world. Writing style of Gurmukhi script is from top to bottom and left to right. In Gurmukhi script, there is no case sensitivity. Presently, fairly printed Gurmukhi script documents and degraded printed Gurmukhi script documents can be recognized by OCR software, but there are very limited efforts in the recognition of complete handwritten Gurmukhi script document. Most of the work on Gurmukhi script recognition system is done by Lehal and Singh [20]. They have developed the complete recognition system for printed Gurmukhi script, where connected components are first segmented using thinning based approach. Algorithm for segmentation of isolated handwritten words was available in 2006 by Sharma and Lehal [21]. They proposed technique for segments the words in an iterative manner by focusing on presence of headline, aspect ratio of characters and vertical and horizontal projection profiles.

Jindal *et al.* [22] have provided a solution for touching character segmentation of printed Gurmukhi script. Also they have provided a very useful solution for segmenting overlapping lines in various Indian scripts [23]. They have proposed the technique for segment the degraded Gurmukhi script into upper, middle, lower zones. They have also provided the complete recognition system for complete degraded printed Gurmukhi script documents [22, 23]. Online handwriting Gurmukhi script recognition system was available in 2008 by Sharma *et al.* [24]. They have used the elastic matching technique in which character is recognized into two stages. In the first stage they recognize the strokes, in the second stage, character is evaluated on the basis of recognized strokes.

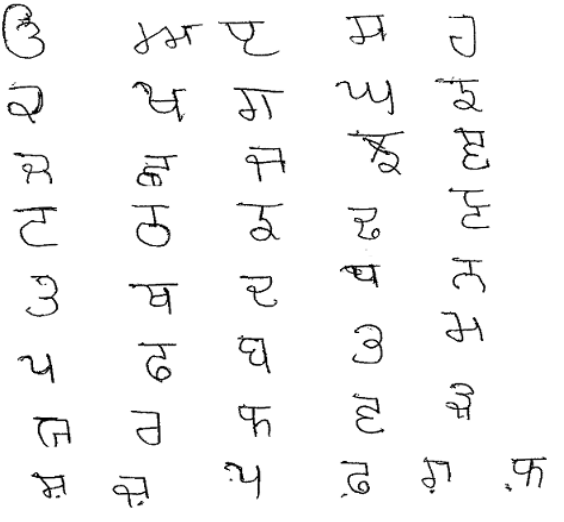

Fig. 6. Handwritten Gurmukhi characters

4 Conclusion and Future Scope

In this paper, we presented work done on handwritten Indian scripts. Firstly, we discussed stages of complete OCR system for handwritten document recognition.

After that we explore the techniques and methods, developed for recognition of particular Indian script. A lot of work has been done on Devanagari and Bangla handwritten characters recognition. Little work has been done to recognize the Oriya, Kannada, Tamil, Malayalam and Gurmukhi handwritten script. But till now there is no complete recognition system is available for recognition of Indian scripts. As such, there is a need for a handwriting OCR for Indian script that can help the people for converting the handwritten text to computer processable format. In future work, the techniques used for Bangla and Devanagari characters can be used for other offline handwritten Indian scripts so that accuracy of recognition can be prefect as Bangla and Devanagri character.

References

1. Pal, U., Chaudhuri, B.B.: Indian Script character recognition: a survey. Pattern Recognition 37, 1887–1899 (2004)
2. Pal, U., Chaudhuri, B.B.: Script line separation from Indian multi –script documents. In: The Proceedings of 5th ICDAR, pp. 406–409 (1999)
3. Pal, U., Wakabayashi, T., Kimura, F.: A system for off-line Oriya handwritten character recognition using curvature feature. In: The Proceedings of 10th ICIT, pp. 227–229 (2007)
4. Chaudhuri, B.B., Pal, U.: An OCR System to Read Two Indian Language Scripts: Bangla and Devanagari (Hindi). In: The Proceedings of 4th ICDAR, vol. 2, pp. 1011–1015 (1997)
5. Bansal, V., Sinha, R.M.K.: Integrating Knowledge Sources in Devanagari Text Recognition. IEEE Transaction on Systems, Man and Cybernetics 30(4), 500–505 (2000)
6. Mori, S., Yamamoto, K., Suen, C.Y.: Historical review of OCR research and development. Proceedings of the IEEE 80(7), 1029–1058 (1992)
7. Kumar, D.: AI approach to hand written Devanagri script recognition. In: The Proceedings of IEEE Region 10th International Conference on EC3-Energy, Computer, Communication and Control Systems, vol. 2, pp. 229–237 (2008)
8. Hanmandlu, M., Murthy, O.V.R., Madasu, V.K.: Fuzzy model based recognition of handwritten Hindi characters. In: The Proceedings of 9th Biennial Conference of the Australian Pattern Recognition Society, pp. 454–461 (2007)
9. Sethi, K., Chatterjee, B.: Machine recognition of constrained hand-printed Devanagari numerals. J. Inst. Elec. Telecom. Engg. 22, 532–535 (1976)
10. Dey, L., Bajaj, R., Chaudhury, S.: Devanagari numeral recognition by combining decision of multiple connectionist classifier. Sadhana 27, 59–72 (2002)
11. Chaudhuri, B.B., Majumder, D.D., Parui, S.K.: A procedure for recognition of connected hand written numerals. Int. J. Systems Sci. 13, 1019–1029 (1982)
12. Pal, U., Roy, K., Kimura, F.: Bangla Handwritten Pin Code String Recognition for Indian Postal Automation. In: The Proceedings of 11th ICFHR, pp. 290–295 (2008)
13. Bishnu, A., Chaudhuri, B.B.: Segmentation of Bangla handwritten text into characters by recursive contour following. In: The Proceedings of 5th ICDAR, pp. 402–405 (1999)
14. Dutta, A., Chaudhury, S.: Bengali alpha-numeric character recognition using curvature features. Patter Recognition 26(12), 1757–1770 (1993)
15. Pal, U., Belaid, A., Choisy, C.: Touching numeral segmentation using water reservoir concept. Elsevier Science Inc. 24(1), 261–272 (2003)
16. Bhattacharya, U., Gupta, B.K., Parui, S.K.: Direction code based features for recognition of online handwritten characters of Bangla. In: The Proceedings of 9th ICDAR, vol. 1, pp. 58–62 (2007)

17. Pal, U., Chaudhuri, B.B.: Skew Angle Detection of Digitized Indian Script Documents. IEEE Transactions on PAMI 19(2), 182–186 (1997)
18. Sundaram, S., Ramakrishnan, A.G.: Two Dimensional Principal Component Analysis for Online Tamil Character Recognition. In: The Proceedings of 11th ICFHR, pp. 88–94 (2008)
19. Aparna, K.H., Subramaniam, V., Kasirajan, M., Prakash, G.V., Chakravarthy, V.S., Madhvanath, S.: Online handwriting recognition for Tamil. In: The Proceedings of 9th International Workshop on Frontiers in Handwriting Recognition (IWFHR), pp. 438–443 (2004)
20. Lehal, G.S., Singh, C.: A Gurmukhi Script Recognition System. In: The Proceedings of 15th International Conference on Pattern Recognition, vol. (2), pp. 557–560 (2000)
21. Sharma, D.V., Lehal, G.S.: An iterative algorithm for segmentation of isolated handwritten words in Gurmukhi script. In: The Proceedings of 18th ICPR, vol. 2, pp. 1022–1025 (2006)
22. Jindal, M.K., Lehal, G.S., Sharma, R.K.: Segmentation Problems and Solutions in Printed Degraded Gurmukhi Script. International Journal of Signal Processing 2(4), 258–267 (2005)
23. Jindal, M.K., Sharma, R.K., Lehal, G.S.: Segmentation of Horizontally Overlapping Lines in Printed Indian Scripts. International Journal of Computational Intelligence Research 3(4), 277–286 (2007)
24. Sharma, A., Kumar, R., Sharma, R.K.: Online handwritten Gurmukhi character recognition using elastic matching. In: The Proceedings of Congress on Image and Signal Processing, vol. 2, pp. 391–396 (2008)
25. Acharya, D., Subba Reddy, N.V., Makkithaya, K.: Multilevel Classifiers in Recognition of Handwritten Kannada Numerals. In: The Proceedings of PWASET, vol. 2, pp. 284–289 (2008)
26. Rajashekararadhya, S.V., Ranjan, P.V.: Efficient Zone Based Feature Extraction Algorithm for Handwritten Numeral Recognition of Four Popular South Indian Scripts. Journal of Theoretical and Applied Information Technology 4(12), 1171–1181 (2008)