International Journal of communication and computer Technologies, ISSN: 2278-9723
Available at http://www.ijccts.org

An Online and Offline Character Recognition Using Image Processing Methods-A Survey

Mr.Ban Maheskumar N [1], Prof.Sayed Akhtar H [2]

[1,2] *M.E Computer Science Department, Aditya Engineering College Beed Maharastra, India,*

Received: 20-08-2016, **Revised:** 15-09-2016, **Accepted:** 21-11-2016, **Published online:** 19-12-2016

***Abstract*— In this paper we are presenting different methods of character recognition that are present in nowadays. Character recognition can be performed by online and offline methods.**

We present a comparative study of all the methods available for the analysis of character recognition that implements image processing methods.

***Keywords*— clustering, context instances , association .**

I. Introduction

In today's world character recognition plays important role to make everything automotive. Optical Character Recognition (OCR) is the process of converting handwritten or printed documented into machine readable form. In today's fast growing technology, digitization of the documents is important for future use which gives scope for the researches to perform Optical Character Recognition.

To perform character recognition ,a document scanner with digitization of document software have been implemented in today's fast growing technology. OCR software allows you to scan a printed document and then convert the electronic text in word format.

OCR receives its attention in the area of digitization in library and digitization of historical documents. In this paper an efficient approach for digitization of the Tamil character have been proposed.

Tamil is one of the accepted language which is currently used by Tamil people. Character recognition in Tamil language is very less because defining of features for the Tamil language is difficult due to its complex character style and large data sets.

Character recognition is a special branch of pattern recognition which refers to the translation of handwritten or printed text into machine readable text.

Offline handwriting recognition system has versatile range of applications including processing of bank chequs, mail addresses, white board reading, recognition of handwritten manuscripts etc.

Handwritten character recognition is divided into:online and offline recognition. The difference originates from the type of input data that is available for recognition .

In online recognition a special input device, e.g. an
electronic pen, tracks the movement of the pen during the writing process. No time information is available in offline handwriting recognition. Only an image of the handwritings processed. Because less information is available, offline recognition is usually considered more difficult and challenging than online recognition.

II. COMPARATIVE STUDY OF CHARACTER RECOGNITION TECHNIQUES

On-line character recognition refers to the process of recognizing handwriting recorded with a digitizer as a time sequence of pen coordinates. In case of online handwritten character recognition, the handwriting is captured and stored in digital form via different means. Usually, a special pen is

used in conjunction with an electronic surface. As the pen moves across the surface, the two-dimensional coordinates of successive points are represented as a function of time and are stored in order . It is generally accepted that the on-line method of recognizing handwritten text has achieved
better results than its off-line counterpart. This may be
attributed to the fact that more information may be captured in the on-line case such as the direction, speed and the order of strokes of the handwriting.
The on-line handwriting recognition problem has a number of distinguishing features which must be exploited to get more accurate results than the online recognition problem

- It is adaptive: The immediate feedback is given by the
writer whose corrections can be used to further train the
recognizer.
- It is a real time process: It captures the temporal or
dynamic information of the writing. This information
consists of the number of pen strokes, the order of pen strokes. The direction of the writing for each pen
stroke and the speed of the writing within each
pen stroke.
- Very little preprocessing is required. The
operations such as smoothing, and feature
extraction operations such as the detection of line
orientations corners loops are easier and faster with
the pen trajectory data than on pixel images.
- Segmentation is easy: Segmentation operations are
facilitated by using the pen lift information particularly
for hand printed characters.
- Ambiguity is minimal: The discrimination between
optically ambiguous characters may be facilitated with
the pen trajectory information Off-line handwriting recognition refers to the process of recognizing words that have been scanned from a surface (such as a sheet of paper) and are stored digitally in grey scale format.

After being stored, it is conventional to perform further processing to allow superior recognition.

The offline character recognition can be further grouped into two types:

- Magnetic Character Recognition (MCR)
- Optical Character Recognition (OCR)

In MCR, the characters are printed with magnetic ink. The reading device can recognize the characters according to the unique magnetic field of each character. MCR is mostly used in banks for check authentication. OCR deals with the recognition of characters acquiring by optical means, typically a scanner or a camera. The characters are in the form of pixelized images, and can be either printed or handwritten, of any size, shape, or orientation. The OCR can be subdivided into handwritten character recognition and printed character recognition. Handwritten Character Recognition is more difficult to implement than printed character recognition due to diverse human
handwriting styles and customs. In printed character recognition, the images to be processed are in the forms of standard fonts like Times New Roman, Arial, Courier, etc

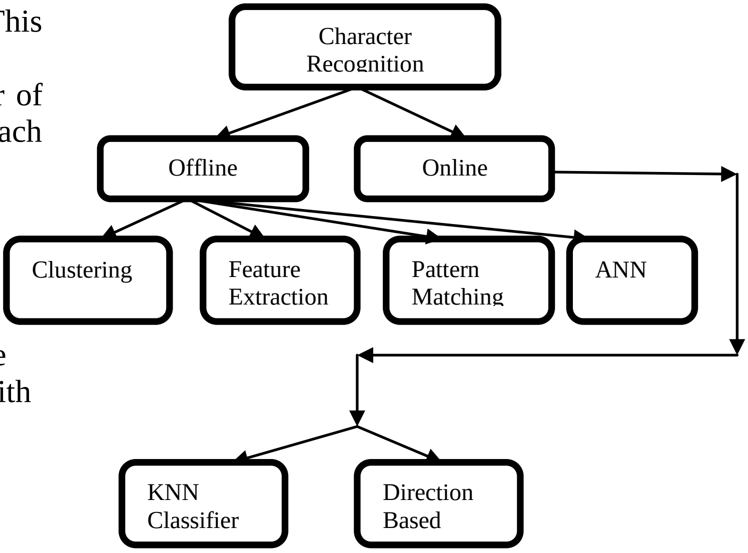

Fig 1.Classification of Character Recognition Techniques

The major steps involved in recognition of characters include, pre processing, segmentation, feature extraction and classification.

1.Pre Processing

a.Noise Removal

b.Thresholding
c.Skeletonizatio
2.Segmentation
3.Feature Extraction
4.Classification
5.Post-processi

Table 1:Comparision of Character recognition techniques

Sr.No.	Comparison	Online Characters	Offline Characters
1	Availability of number of pen strokes	Yes	No
2	Raw Data requirements	Samples/second	Dots/inch
3	Way of writing	Using digital pen on led	Paper document
4	Recognition rates	Higher	lower
5	accuracy	Higher	lower

Character recognition systems extensively use the methodologies of pattern recognition, which allots an
unknown sample to a predefined class. Many techniques
for character recognition are investigated by the researchers and character recognition approaches can be classified as [5] Template matching, Statistical techniques, Syntactic or structural, Neural network, Hybrid or Combination approaches.

- Template matching approach

This is the simplest way of character recognition, based on matching the stored data against the character to be recognized. The matching operation determines the degree of similarity between two vectors i.e. group of pixels, shapes curvature etc. a gray level or binary input character is compared to a standard set of stored data set. According to similarity measure (e.g. Euclidean, Yule similarity measures etc.), a template matcher can ombine multiple information sources, including match strength and k-nearest neighbor measurements from different
matrices. The recognition rate of this method is very sensitive to noise and image deformation.

- Statistical Techniques

Statistical decision theory is concerned with statistical decision functions and a set of optimality criteria, which increases the probability of the observed pattern given the model of a certain class. Statistical techniques are based on the assumptions such as Distribution of the feature set,statistics available for each class, collection of images to extract a set of features which represents each distinct class of patterns. The measurements are taken from n-features of each word unit that can be thought to represent an n-dimensional vector space. The major statisticalmethods applied in the character recognition field are Nearest Neighbor Likelihood
or Bayes classifier, clustering Analysis, Hidden Markov
Modeling, Fuzzy Set Reasoning,Quadratic classifier etc.

- Syntactic or Structural Approach

In Syntactic Pattern recognition a formal analogy is drawn between the structure of pattern and syntax of a language. Structural pattern recognition is intuitively appealing because in addition to classification, this approach also gives adescription of how the given path constructed from the primitives. Flexible structural matching is proposed for identification of alphanumeric characters

- Neural Networks

Various types of neural networks are used for character recognition classification. A neural network is a computing architecture that consists of massively parallel interconnection of adaptive neural processors. Because of its parallel nature, it can perform computations at a higher rate compared to classical
techniques. Because of its adaptive nature, it can adapt to changes in the data and learn the characteristics of input signal. Output from one node is fed to another one in the network and final decision depends on the complex interaction of all nodes. Several approaches exist for training of neural networks like error correction, Boltzman, Hebbian and competitive learning. Neural network architectures can be classified as, feed-forward, feed-back and recurrent networks. The most

common neural networks used in the character recognition systems are the Multi Layer Perceptron (MLP) of the feed forward networks and the Kohonen's Self Organizing Map of the feedback networks.

- Hybrid or Combination Classifier

We may have different classification methods or different training sections, different feature sets, different training sets, all resulting in set of classifiers, whose outputs may be combined together, with the hope of improving overall classification accuracy. If this set of classifiers is fixed, the problem mainly focuses on the combination function. It is possible to use a fixed combiner and optimize the set of input classifiers. A typical combination scheme consists of a set of individual combiner and classifiers which combines the results of the individual classifiers to make the final
decision. Various schemes for combining multiple
classifiers can be grouped into three main categories
according to their architecture cascading, hierarchical, and parallel.

- Indian Character Recognition

Not many attempts have been made on the character recognition of Indian character sets. However, some major works are reported on Devanagari. Some attempts are also reported on Tamil, Kannada, Gujarathi, Bengali, Malayalam and Telugu.Character recognition of handwritten and printed text is of great importance for electronic conversion of historical information including letters, diaries, wills and other manuscripts. The problem is challenging because of human handwriting variability, uneven skew and orientation as well as noise and distortion such as smudges, smears, faded print, etc.
identification of handwritten Indian scripts especially of Bangala, as well as Malayalam, Hindi, English, etc. Most of the Indian scripts have 500 or more characters or symbols used in running text, through the number of basic vowels and consonants is not more than 50. The number is multiplied by three types of vowel modifiers that may be glued below the consonants, thus generating threefold consonant-vowel combinations.

Further increase in number is possible where consonant creates a complex orthographic shape called compound characters. For some scripts like Bangla, Gujarthi, Telugu and Devanagari languages consists of large number of compound characters. These compound characters can also take vowel modifiers to generate threefold more shapes. Thus orthographic shapes may run of the order of thousand. Only Tamil and Punjabi scripts are relatively simpler, where the number of characters/ symbol is about 150 and 70 respectively. Most Indian script lines can be partitioned into three sub-zones. The upper and lower zones may consist of parts of the basic characters as well as vowel modifiers. These parts of two consecutive text lines normally do not overlap or touch in case of printed script, but for handwriting, people have the tendency to write them
bigger, leading to overlapping and touching characters.
Overall these characteristics make handwritten and printed Indian text recognition more challenging

III.APPLICATIONS

Optical Character Recognition has a wide range of
applications in various areas. It can be used as a
telecommunication aid for post al address reading for the
deaf, processing of documents, in recognition of foreign
language and also for language translation [32]. In bill
processing systems it is used to read payment slips like
electricity bills , telephone / water bills. It will read and
recognize the amount to be paid and also recognize the
account number. The character recognition system can also be used for reading the address, assigning Zip codes to letters, application forms, voter ID cards, and identification of bank cheques by recognizing the account number and the amount written on the cheque
. These systems can also be used in automatic processing of issuing tickets to air line passengers, validation of passports and visa cards etc. Address readers in postal departments locates the address on letters and sorts them according to their location

using the zip code. The multiline optical character reader (MLOCR) by United States Post al Services (USPS) locates the address block on a mail piece, reads the address, identifies ZIP and generates a 9-digit bar code and sorts the mail to the correct stacker. This classifier recognizes up to
400 fonts and the system can process up to 45,000 mail
pieces per hour [33].

IV.CONCLUSION

Apparently, digital image processing is an important aspect of photography considering that technology keeps changing. There are a host of digital image processing techniques that provides a wide application variety in feature extraction and classification. Artificial neural networks are frequently used to undertake character recognition because of their high tolerance to noise. The systems have the capability to realize perfect results. Apparently, the feature extraction stage of OCR is the most significant. Survey represents a study of feature extraction methods with different classifiers implemented in OCR systems for different Indian scripts .Variance between the features should be clearly discriminative and specific so that system can classify the characters with maximum efficiency and minimum error rate. This survey paper helps researchers and developers to understand history of the OCR research work for Indian scripts. OCR for Indian scripts that works under all possible conditions and gives highly accurate results still remains a highly challenging task to implement. We believe that our survey will be helpful for researchers in this field.

References

1] G. Tauschek, “Reading machine,” U.S. Patent 2 026 329, Dec. 1935. [23] P. W. Handel, “Statistical ma chine,” US. Patent 1915 993, June 1933.
[2] P. W. Handel, “Statistical machine,” US. Patent 1915 993, June 1933.
[3] R. Plamondon and S. N. Srihar i, “On-line and o ff-line handwritten recognition: a comprehensive survey”, IEEE Transactions on PAMI, Vol. 22(1), pp. 63–84, 2000.
[4] T. Yarman, Nafiz arica, factos – “An overview of CR focused on offline handwriting “– IEEE – 1996. Jangala. Sasi Kiran et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 2065-2069 www.ijcsit.com 2068
[5] K. Anil Jain, “Statistical Pattern Recognition: A Review”, IEEE Trans. Pattern Analysis and Mach ine Intelligence, Vol. 22, 1, 2000, pp. 4-37.
[6] Lim, Jae S.,” Two-Dimensional Signal and Image Processing”, Englewood Cliffs, NJ, Prentice Hall, 1990, pp. 469-476.
[7] P. K. Sahoo, S. Soltani, A.K.C Wong and Y C Chen, “A survey ofThresholding Techniques”, Comput er Vision, Graphics and Image processing, vol 41, pp 233-260, 1988.
[8] Otsu.N, "A threshold selection method from gray level histograms", IEEE Trans. Systems, Man and Cy bernetics, vol.9, pp.62-66, 1979
[9] L. Lam, S.W. Lee and C.Y.Suen, “Thinning Methodologies: A Comprehensive Survey”, IEEE Trans. Pattern Anlaysis and Machine Intelligence, vol.14. pp 869-885,1992
[10] Richard G. Casey And Eric Lecolinet “ A Survey Of Methods And Strategies In Character Segmentation”, IEEE Trans. On Pattern Analysis And Machine Intelligen ce, Vol 18, Pp 690-706,1996
[11] Trier.O.D, Jain.A.K and Taxt.J, "Feature extraction methods for character recognition - A survey", Pattern Recognition, vol.29, no.4, pp.641-662, 1996.
[12] Veena Bansla and R M K Sinha, “A Complete OCR for printed Hindi Text in Devanagari Script”, IEEE 800 - 804 2001.
[13] Sinha. M. K., Mahabala., “Machine Recognition of Devnagari Script”, IEEE T. SYST. MAN Cyb., vol.. 9,pp.435-449,1979.
[14] Reena Bajaj, Lipika Dey and Santanu Chaudhury, “Devnagari numeral recognition by combining de cision of multiple connectionist classifiers”, Vol. 27, Part 1, February 2002, pp. 59–72.
[15] Sandhya Arora, “A Two Stage Classification Approach for Handwritten Devanagari Charact ers”, IEEE 399 - 403 vol 2.
[16] Pritpal Singh and Sumit B udhiraja, “Feature Extraction and Classification Techniques in O. C.R. Systems for Handwritten Gurmukhi Script”. International Jo urnal of Engineering Research and Applications (IJERA), Vol.1, ISSUE 4,pp.1736-1739.
[17] Gita Sinha Rajneesh Rani Renu Dhir , “Handwritten Gurmukhi Numeral Recognition using Zone-based Hybrid Feature Extraction Techniques”, International Journa l of Computer Applications(0975-8887), Volume 47- No. 21 June 2012.
[18] G. S. Lehal and Chandan Singh, “Feature extraction and Classification for OCR of Gurmukhi Script”.
[19] Dharamveer Sharma, Puneet Jhajj, “Recognition of Isolated Handwritten Charactersin Gurmukhi S Volume 4– No.8, August 2010cript”, International Journal of Computer Applications (0975 – 8887)

[20] Jalal Uddin Mahtnud, Moha mmed Feroz Raihan and Chowdhury Mofizur Rahman, “A Complete OCR System for Continuous Bengali Characters”,IEEE 1372 - 1376 Vol. Oct. 2003.
[21] B. B. Chaudhuri and U. Pal, “An OCR System to Read Two Indian Language Scripts: Bangla and Devnagari (Hindi)”, IEEE 1011 - 1015 vol.2, Aug 1997.
[22] U. Bhattacharya1, M. Shridhar, and S.K. Parui1, “On Recognition of Handwritten Bangla Characters”.
[23] Atul Negi, Chakravarthy Bhagvati, B. Krishna, “An OCR system for Telugu”, IEEE 1110 – 1114 -2001.
[24] Vasantha Lakshmi, C. Patvardhan, C. , “A high accuracy OCR system for printed Telugu text”, IEEE 725 - 729 Vol.2 , Oct-2003.
[25] Arun K Pujari, C Dhanunjay Naidu, “An Adaptive Character Recognizer for Telugu Scripts using ultiresolution Analysis and Associative Memory”.
[26] R Sanjeev Kunte, R D Sudha ker Samuel, “An OCR System for Printed Kannada Text Usi ng Two - Stage Multi-network Classification Approach Employing Wavelet Features”, IEEE 349 – 353,Dec-2007.
[27] Sagar, B.M. , Shobha, G. , Kumar, P.R. , “Complete Kannada Optical Character Recognition with syntactical analysis of the script”, IEEE 1 – 4 , Dec. 2008.
[28] T V Ashwin and P S Sastry , “A font and si ze-independent OCR system for printed Kannada doc uments using support vector machines”, Vol. 27, Part 1, February 2002, pp. 35–58.
[29] Chaudhuri, B.B, Pal, U. , Mitra, M. , “Automatic recognition of printed Oriya script”, IEEE 795 – 799, 2001.
[30] Tariq, J, Nauman, U, Naru, M.U , “Softconverter: A novel approach to construct OCR for printed Urdu isolated characters”, IEEE V3-495 - V3-498 ,April 2010.
[31] Sardar, S, Wahab, A, “Optical character recognition system for Urdu”,IEEE 1 - 5 , June 2010.
[32] Aditi Goyal, Kartikay Kha ndelwal, Piyush Keshri “ Optical Character recognition for Handwritten Hindi” Stanford University.
[33] Divya Sharma “ Recognition of Handwritten Devnagari Script using Soft computing”, Thesis report, Master of Engineering, Thapar University.