A Literature Survey on Character Recognition of Indian Scripts for New Researchers

[1]Prof. BambKalpesh K, [2]Prof. Tated K.S,[3] Prof. Mutha H.H., [4]Prof. Chopda P.P.

Department of Electronics and Telecommunication Engineering,
SNJB KBJ'S College of Engineering, Chandwad, Nashik[1234]
kalpeshkumar.bamb@gmail.com[1]

Abstract-

After the printed character recognition handwritten character recognition is always a leading area of research in the field of pattern recognition. Even though, sufficient studies have performed in foreign scripts like Arabic, Chinese and Japanese, only a very few work can be traced for handwritten character recognition mainly for the south Indian, Devanagari scripts. OCR system development for Indian script has many application areas like banking, libraries, preserving manuscripts and ancient literatures written in different Indian scripts and making digital libraries for the documents. Feature extraction and classification are the two essential steps of character recognition process affecting the overall accuracy of the recognition system. This surveyrepresents a history of character recognition with digital image processing techniques such as Feature Extraction, classification, Image Restoration and Image Enhancement.

Keywords:*Optical Character Recognition (OCR), Feature Extraction, Classification, Digital Image Processing.*

INTRODUCTION

The history of OCR can actually found back in 1923 Tausheck [4] and 1933 Handel [2] gave the first idea of the concept of the OCR. Optical Character Recognition deals with the problem of recognizing optically processed characters. Optical recognition is an offline process i.e. the recognition starts after writing or printing has been completed. Handwritten character recognition is a frontier area of research for the past few decades and there is a large demand for OCR on handwritten documents. Even though, sufficient studies have performed in foreign scripts like Chinese, Japanese and Arabic characters [3], only a very few work can be traced for handwritten character recognition of Indian scripts. Even now no complete hand written text recognition system is available in Indian scenario and it is difficult due to large character set of Indian languages and the presence of vowel modifiers and compound characters in Indian script. The problem of character recognition can be classified based on two criteria. One is based on the type of the text which is printed or hand written. The other is based on the acquisition process which can be on-line of off-line. It is generally considered that the on-line method of recognizing handwritten text has achieved better results than its off-line counterpart. In case of online character recognition, there is real time recognition of characters. Online systems havebetter information for doing recognition since they have timing information and can avoid the initial search step of locating the character as in the case of their offline counterpart. In case of offline character recognition, the typewritten or handwritten character is typically scanned in the form of a paper document and made available in the form of a binary or gray scale image to the recognition algorithm. Offline character recognition is a more challenging and difficult task as there is no control over the medium and instrument used.

1. Brief History of Character recognition

Many methods have been proposed for character recognition; they are often subjected to substantial constraints due to unexpected difficulties. Historically character recognition system has evolved in three ages [1], as;

1.1 1895-1975 (initial ages) – The history of characterrecognition can be traced as early as 1900. When the Russian Scientist Tyering attempted to develop an aid for visually handicapped the first character recognizers appeared in the middle of 1940s with the development of digital computers. The previous work on the automatic recognition of characters has been concentrated either upon machine printed text or upon small set of well distinguished hand written text or symbols. The commercial character recognizers were available in 1950s.

1.2 1975-1990 – The studies until 1975suffered from the tack of powerful computer hardware and data acquisition derives. However, the character recognition research was focused on basically the shape recognition techniques without using any semantic information.

1.3 After 1990 till date – The real progress oncharacter recognition system is achieved during this period, using the new

methodologies and development tools, which are empowered by continuously growing information technologies. In the early nineties, Image processing and Pattern recognition techniques are efficiently combined with the Artificial Intelligence methodologies.

These days in addition to the more powerful computers and accurate electronic equipments such as cameras, scanners and electronic tablets, we have efficient use of methodologies such as Hidden Markovmodels, neural networks;Fuzzy set reasoning and Natural language processing. Character recognition system is the base for many different types of applications in various fields, which we use in our daily lives. Post offices, banks, security systems, number plate recognition system and even in the field of robotics use this system as the base of their operations.

CHARACTER RECOGNITION APPROACHES

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.

2.1 Template matching approach - This is the simplestway 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 combine 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.

2.2 Statistical Techniques - Statistical decision theory isconcerned 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 statistical methods applied in the character recognition field are Nearest Neighbor Likelihood or Bayes classifier, clustering Analysis, Hidden Markov Modeling, Fuzzy Set Reasoning, Quadratic classifier etc.

2.3 Syntactic or Structural Approach - In SyntacticPattern 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 a description of how the given path constructed from the primitives. Flexible structural matching is proposed for identification of alphanumeric characters.

2.4 Neural Networks - Various types of neural networksare 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.

2.5 Hybrid or Combination Classifier - We may havedifferent 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.

2.6 Indian Character Recognition - Not many attemptshave been made on the character recognition of Indian character sets. However, some major works are reported on Devanagari. Some attempts are also reported onhindi, Marathi, Tamil, Kannada,

Gujrathi, 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 theconsonants, 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.

ARCHITECTURE OF A GENERAL CHARACTERRECOGNITION SYSTEM

The major steps involved in recognition of characters include, pre processing, segmentation, feature extraction and classification (figure 1.)

3.1 Pre Processing - The sequences of pre-processing stepsare as follows.

3.1.1 Noise Removal - Noise is defined as any degradationin the image due to external disturbance. Quality of handwritten documents depends on various factors including quality of paper, aging of documents, quality of pen, color of ink etc. Some examples of noises are Gaussian noise, salt and pepper noise. These noises can be removed to certain extent using filtering technique. Technical details of filtering can be observed in [6].

3.1.2 Thresholding - The task of thresholding is to extractthe foreground (ink) from the background (paper) [7]. Given a threshold (T) between 0 and 255, replace all the pixels with gray level lower than or equal to T with black (0), the rest with white (1).

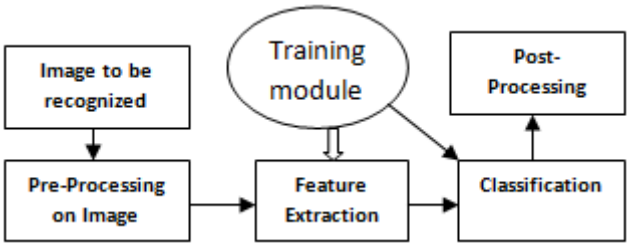

Fig.1Architecture of a character recognition system

If the threshold is too low, it may reduce the number of objects and some objects may not be visible. If it is too high, we may include unwanted background information. The appropriate threshold value chosen can be applied globally or locally. Otsu's [8] algorithm is the commonly used global thresholding algorithm.

3.1.3 Skeletonization - Skeletonization is an imagepreprocessing technique performed to make the image crisper by reducing the binary valued image regions to lines that approximate the skeletons of the region. A complete survey of thinning methodologies is discussed in [9]

3.2 Segmentation - Segmentation step contains wordsegmentation, character segmentation and line segmentation. Methods for character segmentations [10] are based on

i) White space and pitch
ii) Projection analysis and

iii) connected component labeling

3.3 Normalization - It is the process of converting therandom sized image into standard sized image. In this, sizenormalization avoids inter class variation among characters. Bilinear, Bi-cubic interpolation techniques are a few methods for size normalization.

3.4 Feature Extraction - Features are a set of numbers thattake the salient characteristics of the segmented image. There are different feature extraction methods for character recognition [15].

3.5 Classification - The feature vector obtained fromprevious phase is assigned a class label and recognized using unsupervised and supervised method. The data set is divided into training set and test set for each character. Character classifier can be Bayes classifier, nearestneighbour classifier, Radial basis function, Support vector machine, Linear discriminate functions and Neural networks with or without back propagation.

3.6 Post-processing - Post-processing step involvesgrouping of symbols. The process of performing the association of symbols into strings is referred to as grouping.

FEATURE EXTRACTION AND CLASSIFICATION TECHNIQUES

VeenaBansal and R.M.K Sinha [12] presented a complete OCR for printed Hindi text written in Devanagari script. The system used following features: Coverage of the region of the core strip, Vertical bar feature, Horizontal zero crossings, Number of positions of the vertex points, Moments, Structural descriptors of the characters for classification, Tree classifiers are used. Overall accuracy obtained at the character level is 93%. Sinha and Mahabala [13] designed a syntactic pattern analysis system for Devanagari script recognition. The system stores structural descriptors for each symbol of the script. They achieved 90% accuracy. Reena, Lipika and Chaudhury [14] have tried to exploit information about similarity between numerals, Style invariant features and stylistic variations. They presented a approach for recognition of handwritten Devnagari numerals using multiple neural classifiers. SandhyaArora [11] have used Intersection features with Neural Network for Devanagari script and achieved 89.12% accuracy.

Singh and Budhiraja [16] presented an OCR system for handwritten isolated Gurumukhi script using Zoning, Projection histogram, Distance profile features, and Background directional features and used Support Vector Machines (SVM) for classification and thus obtained 95.04% of overall accuracy. Further Geeta and Rani [17] represented an OCR system for Gurumukhi numerals using Zone Distance features and SVM classifier and achieved 99.73% accuracy. G. S. Lehal and Chandan Singh [18] directed their efforts towards development of OCR system for Gurumukhi. They used Local features (concave/convex parts, number of endpoints, branches, joints) and Global features (number of holes, projection profiles, connectivity etc.). For classification hybrid classification technique, binary decision tree and nearest neighbor was used. They achieved a recognition rate of 91.6%. Dharamveer Sharma and PuneetJhajj [19] used zoning feature with hybrid classification technique using KNN and SVM classifier andachieved 72.7% accuracy.

A very influential attempt made by the Jalal, Feroz and Choudhuri [20] for Bangla script. They represent neural network classifier by using Bounded rectangle calculation, Chain code generation, Slope distribution generation features. They achieved 96% system accuracy. Chaudhuri and Paul [21] represent an OCR system to recognize Bangla and Devanagari using stroke and shaded portion feature with tree classifier. U. Bhattacharya, M. Shridhar, and S.K. Paruil [22] implemented Neural network classifier for isolated Bangla characters with chain code features and achieved 92.14% accuracy on testing sets and 94.65% on training sets. Negi and Chakravarthy [23] represent an OCR system with 92% performance using template matching, fringe distance for Telegu script. Another attempt was made by Patvardhan and Lakshmi [24] for Telegu script. They used neural classifier by using directional features and they achieved 92% accuracy. Arun K Pujari, and C Dhanunjaya Naidu [25] implemented an adaptive character recognizer for Telugu scripts using Multi resolution Analysis. They represented DNN (Dynamic Neural Network) using Wavelet analysis and achieved 93.46 % success rate. In south India, Kannada and Telugu have similar scripts.

R Sanjeev and R D Sudhakar [26] represent an OCR system for printed Kannada Script using two stage Multi-Network (Neural Network) classification technique employing wavelet feature and achieved 91% accuracy at character level. M Sagar,

Shobha and Ramakanth [27] designed a syntactical analysis system using Ternary Tree based classification for isolated Kannada characters. They have given more emphasis on Post-processing step, using dictionary based approach to increase the OCR accuracy. T V Ashwin and P S Sastry [28] represents a font and size-independent OCR system for printed Kannada documents using support vector machines (SVM).

B Chaudhuri U Pal and Mitra [33] gave a prototype OCR system for Oriya script. They use Directional features and Global Features and classified them using Decision tree classifier and achieved 96.03% accuracy at character level. Junaid, Umar, and Muhammad Umair [30] attempted to make an OCR system for isolated Urdu characters using NN classifier using structural features like width, height and checksum of the character. Their prototype gained the accuracy of 97.43%. Another good attempt was made by Jhuwair and Abdul [31] for Urdu script. They achieved the 97.12% recognition rate using Sliding window and Hu-moment feature using KNN classifier.
Bamb K K, Zope R.G and Sharma K S [34] gave a new approach which implements combination of template matching and statistical techniques to recognize the devnagari plane script in which they used pixel information for recognition.

APPLICATIONS

Character Recognition has aextensive range of applications in various fields. 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 toletters, application forms, voter ID cards, and identification of bank cheques by recognizing the account number and the amount written on the cheque. It can be used as a telecommunication aid for postal address reading for the deaf, processing of documents, in recognition of foreign language and also for language translation [32]. 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 Postal 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 [29].

CONCLUSION

There are lots of digital image processing techniques that provide 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. it seems that, 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. Discrepancy 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.

REFERENCES

[1] T. Yarman, Nafizarica, factos – "An overview of CR focused on offline handwriting "– IEEE – 1996.
[2] P. W. Handel, "Statistical machine," US. Patent 1915 993, June 1933.
[3] R. Plamondon and S. N. Srihari, "On-line and off-line handwritten recognition: a comprehensive survey", IEEE Transactions on PAMI, Vol. 22(1), pp. 63–84, 2000.
[4] G. Tauschek, "Reading machine," U.S. Patent 2 026 329, Dec. 1935. [23] P. W. Handel, "Statistical machine," US. Patent 1915 993, June 1933.
[5] K. Anil Jain, "Statistical Pattern Recognition: A Review", IEEE Trans. Pattern Analysis and Machine 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", Computer 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 Cybernetics, 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 Intelligence, Vol 18, Pp 690-706,1996
[11] SandhyaArora, "A Two Stage Classification Approach for Handwritten Devanagari Characters", IEEE 399 - 403 vol 2.
[12] VeenaBansla 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, LipikaDey and SantanuChaudhury, "Devnagari numeral recognition by combining decision of multiple connectionist classifiers", Vol. 27, Part 1, February 2002, pp. 59–72.
[15] 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.
[16] Pritpal Singh and SumitBudhiraja, "Feature Extraction and Classification Techniques in O.C.R. Systems for Handwritten Gurmukhi Script". International Journal of Engineering Research and Applications (IJERA), Vol.1, ISSUE 4,pp.1736-1739.
[17] Gita Sinha Rajneesh Rani RenuDhir , "Handwritten Gurmukhi Numeral Recognition using Zone-based Hybrid Feature Extraction Techniques", International Journal of Computer Applications(0975-8887), Vol 47-No.21 June 12.
[18] G. S. Lehal and Chandan Singh, "Feature Extraction and Classification for OCR of Gurmukhi Script".
[19] Dharamveer Sharma, PuneetJhajj, "Recognition of Isolated Handwritten CharactersinGurmukhi S Volume 4– No.8, August 2010cript", International Journal of Computer Applications (0975 – 8887)
[20] Jalal UddinMahtnud, Mohammed FerozRaihan and ChowdhuryMofizurRahman, "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] AtulNegi, ChakravarthyBhagvati, 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 Multiresolution Analysis and Associative Memory".
[26] R SanjeevKunte, R D Sudhaker Samuel, "An OCR System for Printed Kannada Text Using 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 size-independent OCR system for printed Kannada documents using support vector machines", Vol. 27, Part 1, February 2002, pp. 35–58.
[29] Divya Sharma " Recognition of Handwritten Devnagari Script using Soft computing", Thesis report, Master of Engineering, Thapar University.
[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] AditiGoyal, KartikayKhandelwal, PiyushKeshri" Optical Character recognition for Handwritten Hindi" Stanford University.
[33] Chaudhuri, B.B, Pal, U. ,Mitra, M. , "Automatic recognition of printed Oriya script", IEEE 795 – 799, 2001.
[34] Bamb K.K, Zope R.G, Sharma K.S, "Devanagari Script Recognition System using Combined Approach", CTSP-STM, Vol-4, Issue-4, p4-10, 2014.
[35]

