iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** IV **Month of publication:** April 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Optical Character Recognition

Dalbir[#1], Sanjiv Kumar Singh[#2]
M.Tech[#1], Assistant Professor#2
Department of Computer Science, Shri Balwant Institute of Science &Technology, Dcrust University Murthal

Abstract - Character recognition is one of the most interesting and challenging research areas in the field of Image processing. English character recognition has been extensively studied in the last half century. Nowadays different methodologies are in widespread use for character recognition. Document verification, digital library, reading bank deposit slips, reading postal addresses, extracting information from cheques, data entry, applications for credit cards, health insurance, loans, tax forms etc. are application areas of digital document processing. This paper gives an overview of research work carried out for recognition of hand written English letters. In Hand written text there is no constraint on the writing style. Hand written letters are difficult to recognize due to diverse human handwriting style, variation in angle, size and shape of letters. Various approaches of hand written character recognition are discussed here along with their performance.
Keywords: Neural Network, Feature extraction, Classification, OCR, On-line Character Recognition.

I. INTRODUCTION

Optical Character recognition has been a subject of research. Pattern recognition has three main steps: observation, pattern segmentation, and pattern classification. Optical Character Recognition (OCR) systems is transforming large amount of documents, either printed alphabet or handwritten into machine encoded text without any transformation, noise, resolution variations and other factors.

In general, handwriting recognition is classified into two types as off-line and On-line character recognition. Off-line handwriting recognition involves automatic conversion of text into an image into letter codes which are usable within computer and text-processing applications. Off-line handwriting recognition is more difficult, as different people have different handwriting styles. But, in the on-line system, On-line character recognition deals with a data stream which comes from a transducer while the user is writing. The typical hardware to collect data is a digitizing tablet which is electromagnetic or pressure sensitive.

When the user writes on the tablet, the successive movements of the pen are transformed to a series of electronic signal which is memorized and analyzed by the computer.

Optical Character Recognition (OCR) is a field of research in pattern recognition, artificial intelligence and machine Vision, signal processing. Optical character recognition (OCR) is usually referred to as an off-line character recognition process to mean that the system scans and recognizes static images of the characters. It refers to the mechanical or electronic translation of images of handwritten character or printed text into machine code without any variation.

II. LITERATURE REVIEW

This section gives a review of Handwritten Character Recognition in different languages. A brief sketch of preprocessing, feature extraction and classification methods for several works are summarized below

A. Global thresholding is used to convert the input image in to bi-level form. Erosion, dilation and thinning are the morphological operations. To obtain the skeleton thinning is used. Line segmentation, word segmentation and character segmentation used for segmenting input image in to individual characters. After all these preprocessing the character image is given to the feature extraction phase for extracting relevant features. Contour-let transform is the Feature extraction method. Total 32 features are obtained by four level decomposition of Contour-let, aspect ratio, ratio of horizontal and vertical grid values. The extracted features are classified by feed forward neural network. Using the 32 features yield 97.3% accuracy.

B. Hassiba Nemmour et al [2] had taken care of handwritten Arabic word recognition. Two approaches for word recognition such as analytic and holistic approaches are explained. Ridgelet transform is used for feature extraction. Advantage of using ridgelet is to highlight the line singularities in the handwritten words. SVM is used as the classifier. This work gives 84% efficiency.

C. Bindu S Moni et al [3] presented Malayalam character recognition based on modified quadratic classifier and directional

features. The preprocessing method used is the size normalization to convert the image in to 72 X 72. To decompose the character image in to blocks or zones meshing technique is introduced. The normalized image is taken for feature extraction and directional features are obtained by applying sobel mask. MQDF is the classifier used. Recognition efficiency of 95.32% is obtained.

D. Nusaibath C et al [4] in offline handwritten character recognition the input image is acquired by digital camera and scanner. Preprocessing is done to correct the input image, binarization and skeletonization is used for this purpose. Line segmentation, word segmentation and character segmentation are used to segment the characters. Most relevant features are identified by Gabor filtering. Additional features taken are aspect ratio, ratio of horizontal and vertical grid values. Neural network is used for classification. Recognition efficiency obtained is 96.80%.

E. Pritpal Sigh et al [5] presented handwritten character recognition for Gurumukhi numerals. Here the stages of character recognition include preprocessing, feature extraction and classification. The filtered image binarised and normalized to 32 X 64. Features are taken as the wavelet coefficients. In addition to this wavelet feature zonal features are also consider here. Zonal density is obtained by dividing the normalized image in to 16 zones. Aspect ratio is the final feature value for making the feature vector. Multi layer neural network that uses back propagation algorithm as the classifier. Recognition rate of 88.8% is obtained.

F. G Y Chen et al [6] suggested invariant pattern recognition. Invariant means features must be independent. Features that are invariant under translation, rotation, scaling is obtained by Fourier transform. For palm print classification Contourlet transform extract the features and invariant feature is obtained by taking Fourier of the Contourlet coefficients. Classifier used is the adaptive classifier Adaboost. For handwritten numerals also Fourier is applied to Contourlet coefficients and invariant features are classified using Adaboost.

G. Mamatha H R et al [7] is the recognition of Kannada numerals. Kannada is the official language of state Karnataka. It is derived from the southern Bramhi lipi. To preprocess the image binarisation and thinning done. To overcome the limitations of wavelet a new approach is introduced called curvelet transform. So the features are curvelet coefficients and standard deviation is the dimension reductiontechnique. For the classification of characters KNN used. This paper gives accuracy of 90.5%.

H. Angshul Majumdar et al [8] focus on the recognition of Bangali characters. Curvelet coefficients are taken as the feature values. Here two thinned and thickened version of the image is considered. The fundamental concept is that if the character can't recognize with original image it will be recognized with morphologically altered variations. For testing and training five variations of input image is taken and classified using KNN. Overall accuracy of 96.80% is obtained.

I. Jomy John et al [9] propose a handwritten character recognition system for Malayalam language. Gradient and curvature are calculated in feature extraction. Directional information from the arc tangent of gradient is used as gradient feature. Strength of gradient in curvature direction is used as the curvature feature. It uses a combination of gradient and curvature feature in reduced dimension as the feature vector. Support Vector Machine used as classifier.

J. Sukhpreet Singh et al[10] propose Handwritten Gurmukhi character recognition for isolated characters. The preprocessing stage reduces noise and distortion, removes skewness and performs skeletonization of the image. Word segmentation and character segmentation is used as segmentation stages. Gabor Filter based methods are used for feature extraction. The extracted features are given to SVM for classification and achieves accuracy of 94.29%

K. Alvaro Gonzalez et al[11] presents an easy and fast method to recognize individual characters in images of natural scenes that is applied after locating text on such images. Feature can be extracted by using Gradient feature. The recognition is based on a gradient direction feature. KNN for clasification. The efficiency of this work is 85.8%.

L. Ashutosh Aggarwal et al[12] propose a method for Isolated Handwritten Devanagari Character Recognition. Binarization, noise removal and skeletonization are used as Preprocessing steps. Thresholding is used as binarization, median filter used for salt and pepper noise removal. Horizontal and vertical segmentations are used to select individual character. Gradient methods are used as Feature Extraction. The obtained feature is passed to SVM for classification. It gives the recognition efficiency of 94%.

M. Karanbir Kaur et al [13] had taken care of English handwritten character recognition. Introduce image cropping, gray the image and binarization as preprocessing steps. 40-point feature extraction is introduced for extracting the features of the handwritten alphabets. ANN is used for classification.

N. Anita Pal et al [14] focus on the recognition of English handwritten character using neural network. Input character is acquired by scanning. Skeletonization and normalization are used as preprocessing steps. The features are extracted from the handwritten character by using boundary tracing along with Fourier Descriptor. Multilayer Perceptron Network is used for the classification of extracted feature.

O. Abdul Rahiman M et al[15] introduces Malayalam handwritten character recognition by using vertical and horizontal line

positional analyzer algorithm. Image were acquired by scanning is converted to gray scale. Noise removal, skeletonization are used for preprocessing. Line segmentation and character segmentation are used to separate individual character. Count and position of vertical and horizontal lines are taken as relevant features. The characters are classified based on the extracted feature using decision tree.

P. Sumedha B. Hallale et al [16] had taken care of directional feature extraction for handwritten character recognition. Noise removal, skeletonization and normalization are used as preprocessing to prepare the input image. Directional features are extracted using sobel mask. Relevant feature extracted are given to neural network for classification.

Q. VedPrakash Agnihotri et al [17] Presented zone based features for the recognition of Devanagari script. Enhance the input image using binarization, noise removal, edge detection and dilation. For getting the zonal feature divide the enhanced image in to 54 equal zones. So 54 features are obtained from each character. The extracted features are classified using genetic algorithm.

III. Phases of General Character Recognition System

A. *Digitization*

Digitization is the process of converting a paper-based handwritten document into electronic format. Here, each document consists of only one character. The electronic conversion is accomplished by using a method whereby a document is scanned and an electronic representation of the original document as a image file format is produced. We used various scanner for digitization, and the digital image was go for next step that is preprocessing phase.

B. *Pre-processing*

In The pre-processing phase, there is a series of operations performed on the scanned input image. It enhances the image rendering it suitable for segmentation the gray-level character image is normalized into a window sized. After noise reduction, we produced a bitmap image. Then, the bitmap image was transformed into a thinned image.

C. *Segmentation*

The Segmentation phase is the most important process. Segmentation is done by separation from the individual characters of an image. Segmentation of handwritten characters into different zones (upper, middle and lower zone) and characters is more difficult than that of printed documents that are in standard form. This is mainly because of variability in paragraph, words of line and characters of a word, skew, slant, size and curved. Sometimes components of two adjacent characters may be touched or overlapped and this situation creates difficulties in the segmentation task. Touching or overlapping problem occurs frequently because of modified characters in upper-zone and lower-zone. Segmentation is an important stage.

D. *Feature Extraction*

In this phase, features of individual character are extracted. The performance of an each character recognition system that depends on the features that are extracted. The extracted features from input character should allow classification of a character in a unique way. We used diagonal features, intersection and open end points features, transition features, zoning features, directional features, parabola curve fitting–based features, and power curve fitting–based features in order to find the feature set for a given character.

E. *Classification and Post Processing*

The classification is the process of identifying each character and assigning to it the correct character class. The classification techniques can be categorized as:

The various classical techniques are Template matching, Statistical techniques, Structural techniques. Whereas the various soft computing techniques are neural networks, Fuzzy logic, Evolutionary computing techniques. The geometric features extracted like dot, line, curve or loops are given as input to the input layer. Each component of the segmented representation is classified as a dot, line, curve, or loop. In each case, the characteristics of the component are determined: if a line, what are its orientation and its size relative to the character frame - short, medium or long. One input neuron is used to encode each of these possible choices (short/medium/long) and each of four possible orientations for a line. One input neuron is used to encode the characteristics of each component extracted by geometric feature extraction technique. Neuron has two modes of operations as training mode and testing mode. In the training mode, the neuron can be trained to fire (or not), for particular input patterns. Post-processing mainly consists of two tasks – output string generation and error detection/correction. Output string generation will reassemble the strings which have been separated in the process of segmentation whereas error detection/correction will

correct errors with the help of dictionary

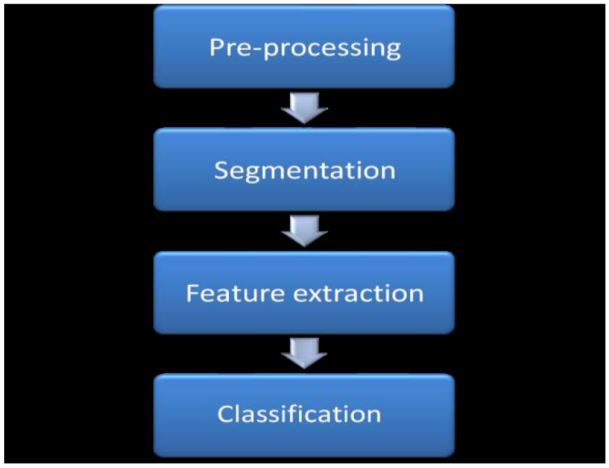


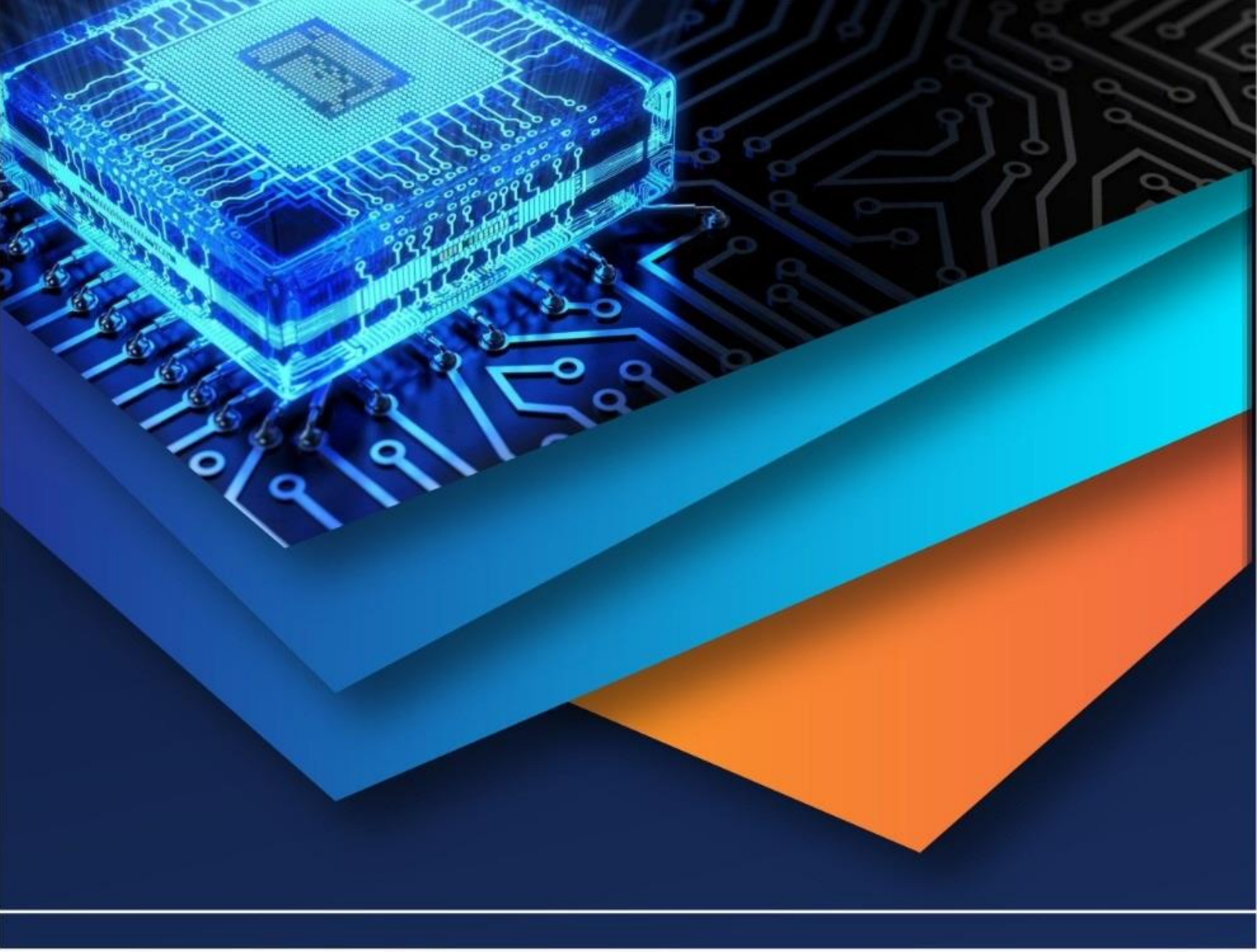
Fig 1.Major Steps of an OCR System

IV. CONCLUSION

The major approaches used in the field of handwritten character recognition during the last decade have been reviewed in this paper. Different pre-processing, segmentation, feature extraction, classification techniques are also discussed. Though, various methods for treating the problem of hand written English letters have developed in last two decades, still a lot of research is needed so that a viable software solution can be made available. The existing OCR for handwritten has very low accuracy. We need an efficient solution to solve this problem so that overall performance can be increased.

REFERENCES

[1]Aji George and Faibin Gafoor,"Contourlet Transform Based Feature Extraction For Handwritten Malayalam Character Recognition Using Neural Network", IRF International Conference Chennai, pp: 107-110, 2014.
[2] Hassiba Nemmour and Youcef Chibani,"Handwritten Arabic Word Recognition based on Ridgelet Transform and support Vector Machines", IEEE, pp: 357-361, 2011.
[3] Bindu S Moni and G Raju,"Modified Quadratic Classifier and Directional Features for Handwritten Malayalam Character Recognition", IJCA Special Issue on"Computer Science-New Dimensions and Perspectives", pp: 30-34, 2011.
[4]Nusaibath C and Ameera Mol P M,"Off-line Handwritten Malayalam Character Recognition using Gabor Filters", International Journal of Computer Trends and Technology, pp: 2476-2479, 2013.
[5]Pritpal Singh and Sumit Budhiraja,"Offline Handwritten Gurmukhi Numeral Recognition using Wavelet Transforms", I. J Modern Education and Computer Science, pp: 34-39, 2012.
[6]G. Y. Chen and B. Kegl,"Invarient Pattern Recognition using Contourlets and Adaboost", Pattern Recognition Society Elsevier, pp: 1-13, 2012.
[7]Mamatha H. R, Sucharitha S and Srikanta Murthy K,"Handwritten Kannada Numeral Recognition based on the Curvelets and Standard deviation", International Journal of Processing Systems, pp: 74-78, 2013.
[8]Angshul Majumdar,"Bangala Basic Character Recognition Using Digital Curvelet Transform", Journal of Pattern Recognition Research, pp: 17-26, 2007.
[9]Jomy John, Kannan Balakrishnan and Pramod K. V, "A system for Offline Recognition of Handwritten Characters in Malayalam Script", I.J.IGSP, pp:53-59, 2013.
[10] Sukhpreet Singh, Ashutosh Aggarwal and Renu Dhir,"Use of Gabor Filters for Recognition of Handwritten Gurmukhi Character", International Journal of Advanced Research in Computer Science and Software Engineering, pp: 234-240, 2012.
[11] Alvaro Gonzalez, Luis M. Bergasa, J. Javier Yebes and Sebastian Bronte"A Character Recognition Method in Natural Scene Images", Pattern Recognition (ICPR), pp: 621-624, 2012.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY