IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

"Comparative Study of Feature Extraction and Classification Techniques for Handwritten Devanagari Script"

Puja Ujwal Talole[1], P. E. Ajmire[2]

M.Sc. II (Computer Science) Dept. Of Computer Science, G. S. Sci., Arts & Comm. College, Khamgaon (MS), India[1]

Head, Dept. of Computer Science, G. S. Sci., Arts & Comm. College, Khamgaon (MS), India[2]

Abstract: The Indian Government has recognized Hindi and English both as official languages, under the VIII scheduled, along with 22 languages. Most of the optical character recognition research work has been done on Devanagari, Telgu, Arabic, and Bangla script. The selection of feature extraction and classification technique is the important to achieve the best accuracy of any recognition system. Features collect the data about the character and accordingly classifier classify the character uniquely. This paper, deals with the studied of the Feature Extraction and Classification techniques for the handwritten Devanagari script.

Keywords: OCR, Feature Extraction, segmentation and Classification techniques.

1. INTRODUCTION

The first idea of optical character recognition was given in 1933. Using this technique computer can recognize characters and other symbols in natural handwritten recognition system [1].Using optical character recognition, the problems are solve for recognition optically processed characters.OCR is an offline process for character recognition, i.e. recognition starts after printing or writing has been completed. For this recognition process the performance is directly dependent on their input parameters, which includes the quality of the image [2].Out of around 33 different languages and 2000 dialects spoken in India, two languages are official and 20 are recognized as scheduled languages for achieving maximum accuracy and performance and minimum error rate in recognition process classification techniques is used. OCR system campers the following steps [3] [4] [5].

2. OPTICAL CHARACTER RECOGNITION STEPS

2.1 Image acquisition process:-

It is the process of document into an electronic form. This has been done with the help of scanning process. With the help of this process digital image of the document is captured. Generally the image is used in black and white form with any format such as JPEG, BMT, and BMP etc. this image is forward to the subsequent blocks for further processing. Image acquisition is the creation of digital images. Digitization produces the digital image which is fed to the pre-processing phase.

2.2 Pre-processing:-

Digital image obtain from scanning process may contain some amount of noise (error), depending of the quality of the scanner.

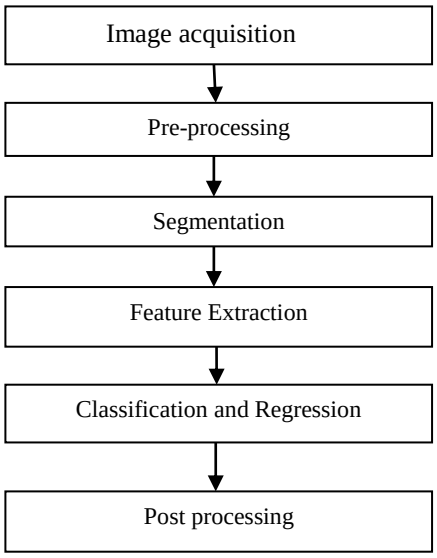

Fig.1: Block Diagram of OCR Steps.

Some of these defects may later cause poor recognition rate [6] and then pre-processing is required to eliminate the noise, Binarization and segmentation [7]. Pre-processing [8] is the initial stage of character recognition. The main steps of pre-processing [9] are as shown in following figure.

2.2.1 RGB to gray scale conversation:-

The scanned image is stored in RGB format. Then this image is converted into gray scale image. It can be represents an image as matrix. Where every element has value of dark and white pixel at the corresponding position should be colored.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

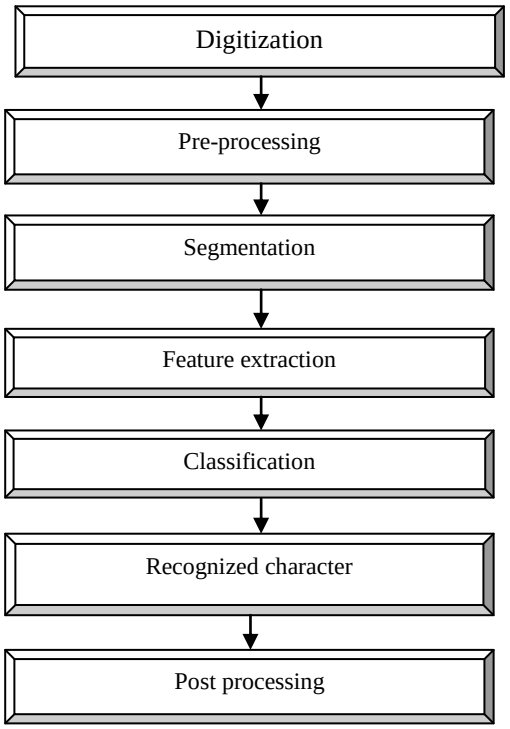

Fig.2: Character recognition system.

2.2.2Binarization:-
In this process gray scale image converted into a binary image using thresholding technique. Thresholding techniques increase the recognition rate of character. Binarization converts all character images to similar form. The processing of binary image also increase the speed of processing as the binary image has only two values i.e. 0 & 1 [10].

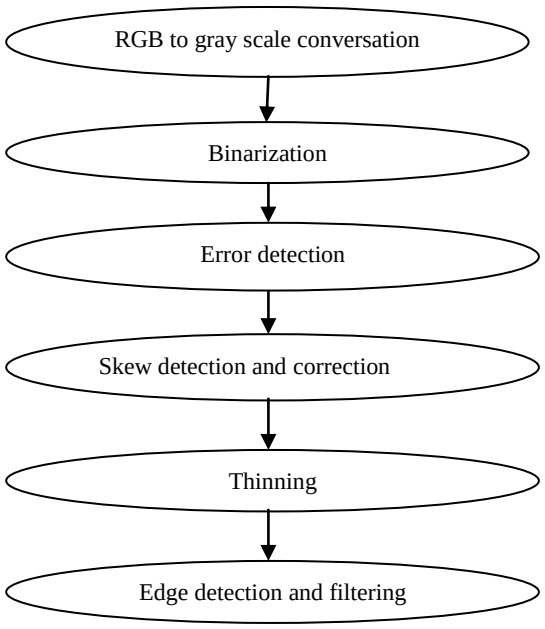

Fig. 3**:** Pre-processing stages.

2.2.3 Error reduction:-
Error reduction is introduces by the optical scanning device which causes the gaps and bumps in lines. It is necessary to eliminate the imperfection. It can be categorized as filtering, morphological operation and noise modeling.

i) Filtering: - This stage causes due to the sampling rate of the digitization stage. The basic goal is to allocate the value to a pixel as a function of the gray values of the neighboring pixels. Filters can be design for sharpening, smoothing, thresholding, or colored background and contrast adjustment purpose.
ii) Morphological operation**:**-The basic task of the morphological operation is to filter the document image. Various morphological operations can be designed to connect the broken strokes, decay the joined strokes, then the characters and extract the boundaries. Hence morphological operations successfully remove noise from the document images due to poor quality of ink and document.
iii) Noise modeling: - Noise can be removed by various calibration techniques, if the noise is in image. Noise modeling is not feasible in most of the application. This removes unwanted data in image.

2.2.4 Skew detection and correction:-
It can specify the deviation of text lines from horizontal or vertical axis. This is caused when the paper is not fed straight in to the scanner. Skew lines are made horizontal by calculating skew angle and making proper correction in the raw image.

2.2.5 Thinning:-
Thinning extract the shape information of the characters. This is the morphological operation which is used to remove selected foreground pixels so that their counters are brought out more widely [10].

2.2.6 Edge detection, dilation and filtering:-
Detection of image in the binaries image is done using sable techniques. After detection the image is dilated and the holes present in the image are filled.

2.3 Segmentation:-
Segmentation is the most important stage in character recognition process/system in Devanagari script because its huge character set and large number of compound characters available in Devanagari text. Using segmentation process the error rate can be minimized. Presence of skewed characters and overlapped causes difficulties in the process of segmentation. It can perform in the following steps [11].

1. Text is segmented into lines.
2. Lines are segmented into words.
3. Words are segmented into characters.

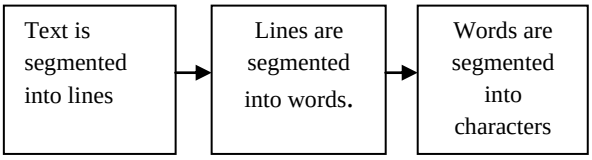

Fig. 4: Segmentation steps.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

Segmentation is of two types are as follows:-

1) External segmentation:-

It is the separation of various writing units such as paragraph, sentences, or words.

2) Internal segmentation:-

In this type, of segmentation the image of series of characters is decomposed into sub-images of individual characters. There are three basic techniques of segmentation [10].

a) The classic approach.
b) Recognition based segmentation.
c) Holistic methods.

2.4 Feature extraction:-

In this stage of character recognition, the essential characteristics of the symbols are captured [2]. It is most important stage. Using this stage redundancy of the data can remove easily. There are two important problems must clarified which are feature extraction and feature selection [12].According to C.Y. Suen. [13], features are obtained from the arrangement of points constituting the character matrix. These features are easily detected as compared to topological features. It can't affect by noise as compared to topological features. It provides low complexity and high speed [14]. A number of techniques are used in feature extraction some are of them as: moment, zoning, projection histogram, n-topples, crossing and distance.

1) Moment: - In this case the moments of different points present in a character are utilized as a feature, S. S. Reddi [15]. Redial and angular moment where as generic moments where proposed by Teh and Chin. [16].
2) Zoning: - The frame of character is divided into several overlapping and no overlapping zones and the densities of object pixels in each zone are calculated. And the density is calculated as a no of object pixel in each zone divided by total no of pixels [17].
3) Projection histogram: - projection histogram gives a no of black pixels in the vertical and horizontal directions of the specified character. It may be vertical, horizontal and left diagonal or right diagonal [18].
4) N-topple: - In this type the position of black and white pixels is considered as feature. This method provides the no of important properties of the pixels.
5) Crossing and distances: - In this type, features are obtained from counts the character image is crossed by vector in certain directions or at certain angles.

2.4.2 Geometrical and topological features:-

This type of represented takes place an encode some knowledge about the stricter of the object or sorting of components make up that object. Various topological and geometrical representations can be grouped into four categories.

1) Extraction and counting topological structure:- In this category, lines, curve, spleens, extreme points, maxima and minima, cups above and below the threshold, opening, and cross points, end lines, loops, direction of a stroke from a special point, inflection between two points, horizontal curve at top or bottom, ascending, descending, and middle stroke and relation among the stroke that make up a character as consider as a feature. [18]
2) Measuring and approximating the geometrical proprieties:- In this type the characters are represented by the measurement of geometrical quantities such as ratio between height and width of the bounding box of character, the relative distance between last point and last y-min, distance between two points, upper and lower masses of words and word length curvature or change in curvature. [18]
3) Coding: - one of the most popular coding is Freeman's chain code. This code is determined by mapping strokes of character into a two dimensional.
4) Graphs and trees: - words or characters are first portioned into a set of topological primitives such as strokes, holes, cross points etc. These primitives are represented using attributed or relational graphs. Trees can also be represented using the words or characters with a set of features, which has a hierarchical relation.

2.5 Classification

Classification is the decision making stage of handwritten character recognition with homogeneous characteristics. It can carried out on the basis of stored features of the features space, i.e. structural and global features etc. classification depends upon quality of features extracted. [20]

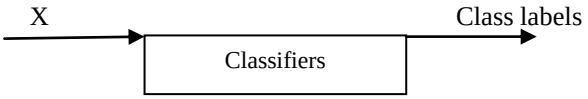

Fig. 5**:** Classifier technique.

Where X is the feature vector.

Some classification methods are temple matching, Statistical technique, Structural techniques, and artificial neural network. [21]

A. Template matching: - This is the simplest form pattern recognition. The given pattern that is to be recognized is compared with the stored pattern. Style and size are ignored while matching the pattern.

B. Statistical method: -In this task, can determined the category of given pattern belongs. By making observations and measurement process, a set of numbers is prepared, which is to prepare a measurement vector. A statistical technique for character recognition is searching of statistical characteristics of various characters [22].

C. Structural method: - It is good for determine handwritten text. In this type, classifies the input pattern on the basis of components of the relationships among these components. Generally a character is represented as production rule structure whose L.H.S. represents character label and R.H.S. represents string of primitives.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

The advantage of the structural approach is that it provides a good symbolic description of the image; however, this feature is more useful for synthesis than analysis tasks [23].

D. Artificial neural network: - It component of inter connected element called neurons. This algorithm is non algorithmic and trainable. The neural network is for pattern classification class is the feed forward network and other used for classification purpose or conventional neural network but the limitation of the system based on neural network is their poor capability for generality.

E. Kernel method: - Most important kernel methods are support vector machine, kernel principle component analysis etc. support vector machine are a group of supervised learning methods that can be applied into training and testing test. Different types of kernel functions of SVM are linier kernel, polynomial kernel, Gaussian radial basis function and sigmoid.

2.6 Post processing

It is the final stage of the recognition system, which involves grouping of symbols. The process of performing the association of the symbols into string is referred to as grouping.

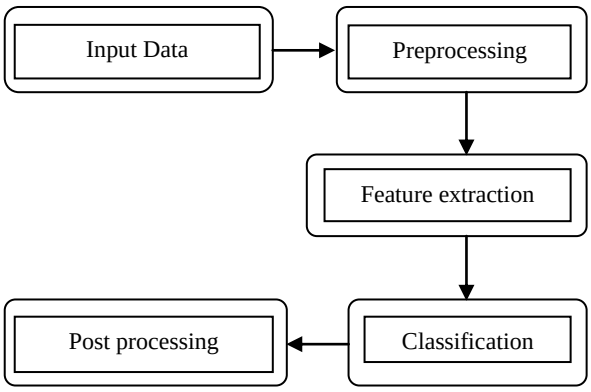

Fig. 6: Post processing steps.

3. CONCLUSION

Survey represents a study of feature extraction methods with different classifiers implemented in OCR system for Devanagari script.

From the table, it is found that the chain code feature extraction and quadratic classifier gives the best recognition accuracy for Devanagari Numerals i.e. 98.86% where as the average recognition rate for all characters such as Vowels, Consonant, Modifier and Compound charter is 94.02%. The following table shows the feature extraction method and classifier which results in best accuracy. This survey paper helps researches and developers to understand history of the research work for Devanagari script. Comparative recognition rate for feature extraction methods are given in following table.

TABLE 1:

Sr. no	Script	Feature Extraction methods	Classifiers.	Recognition rate (%)
1	Hand Written Devanagari	Gradient Features	SVM	94%
2	Handwritten Devanagari	Histogram Gradient	SVM	92.11%
3	Handwritten Devanagari	Rectangle Histogram Oriented Gradient	SVM	95.64%
4	Handwritten Devanagari Modifier	Shape Features	SVM	88.88%
5	Handwritten Devanagari Compound Character	7^{th} order Central Moment	SVM	93.87%
6	Handwritten Devanagari	Chain Code, Zone Based Centroid	SVM ANN	97.61%
7	Handwritten Devanagari	shadow feature and chain code histogram	MLP	92..16 %
8	Handwritten Devanagari	Rectangle Histograms Oriented Gradient	FFANN	97.15%
9	Handwritten Devanagari Numeral	chain code features	Quadratic Classifier	98.86%
10	Devanagari	Syntactical analysis	Tree classifier	90%

REFERENCES

[1] Aayush Purohit and Shardul Singh Chauhan, " A literature Survey on Handwritten Character Recognition", International Journal of Computer applications & Information Technology, Vol.7 (1), 2016.
[2] Er. Harpreet Kaur, "A Survey of Feature Extraction and Classification Techniques Used In Character Recognition for Indian Scripts," International Journal of Engineering Sciences, Vol. 3, Issue December 014.
[3] R Plamndon and S.N.Shrihari, "On-line and off-line handwritten recognition: a compressive survey," IEEE Trans on PAMI, vol.22 (1), PP.62-84, 2000.
[4] R.M.K. Sinha, "Rule based contextual Post Processing for Devangari Text Recognition", Pattern Recognition, Vol.20, NO. 5, pp.475-485, 1987.
[5] R.M.K. Sinha, "On Partining a Dictionary for Visual Text Recognition", Pattern Recognition, Volume 23, pages 497-500 Issue 5, 1990,
[6] K. Prasad, D. Nigam, A. Lakhotiya and D. Umre, " Character Recognition Using mat lab's Neural Network Toolbox", International Journal of u-and e-Service, Science and Technology Vol. 6, no. 1, 2013.
[7] P. M .Kakde and S.M. Gulhane, "Handwritten Devangari Script Recognition: A Review", IJETAE, ISSN 2250-2459, ISO 9001:2008 certified Journal, Vol.4 issue 9, September 2014.
[8] Vyas, M. Verma, K. A."Compressive Survey of Character Segmentation" IEEE, 2014.

[9] Veena Bansal and R.M.K. Sinha, "Integrating Knowledge Source in Devangari Text Recognition System," Technical Report, TRCS-97-248, I.I.T. Kanpur, India, 1997.
[10] Sonal Khare and Jaiveer Singh, "Handwritten Devangari Character Recognition System: A Review," International Journal of Computer Applications (0975-8887) volume. 121, July 2015.
[11] N. K. Garge, Dr. L. Kaur and Dr. M. Jindal, "Recognition of Off-line Handwritten Hindi Text using SVM", International Journal of Image Processing (IJIP), Vol. 7, issue 4, 2013.
[12] Muhammad 'Aarif Mohamed , Dewi Nasien, Haswadi Hassan and Habibollah Haron, "A Review on Feature extraction and Feature Selection for Handwritten Character Recognition", International Journal of Computer Science and Applications, Vol. 6, 2015.
[13] C.Y. Suen, "Character recognition by computer and applications", in Handbook of pattern recognition and Image processing, New York: Academic, pp. 569-586, 1986.
[14] Priyanka Varma and Pratibha Singh, "A survey on handwritten character recognition using Artificial Neural Network," International Journal of Engineering applied sciences and Technology Vol. 1, Issue 5, 2016.
[15] S.S. Reddi, "Radial and angular moment invariants for image identification", IEEE Transactions on PAMI, Vol. 3(2), pp.240-242, 1981.
[16] C. H. Teh and R.T. Chin, "On image analysis by the method of moments", IEEE Transactions on PAMI, Vol. 10(4), pp.496-513, 1988.
[17] Pritpal Singh and Sumit Budhiraja, "Feature Extraction and Classification Techniques in O.C.R. System for Handwritten Gurumukhi Script- A Survey," International Journal of Engineering Research and Applications, Vol. 1, Issue 4,pp. 1736-1739, August 2012.
[18] C.Y. Suen, "Character recognition by computer and applications", in Handbook of Pattern Recognition and Image Processing, New York: Academic, pp. 569-586, 1986.
[19] Snehal S. Patwardhan and R. R. Deshmukh, "A Review of Offline Handwritten Recognition of Devangari Script," International Journal of Computer applications(0975-8887) Volume 117-No. 19, May 2015.
[20] Ratnashil N Khobragade and Dr. Nitin A koli, "A survey on recognition of Devnagari Script," International Journal of Computer applications & Information Technology Vol. II, Issue I, January2013.
[21] P.E.Ajmire, Rajeev Dharaskar, V. M. Thakare, "Statistical Techniques for Feature Extraction for HandwrittenCharacter Recognition"- A Survey, January2014.
[22] P.E.Ajmire, Rajeev Dharaskar, V M Thakare and S E Warkhede, "Structural Features for Character Recognition System-A Review", International Journal of Advanced Research in Computer Science, Volume 3, No. 3, May-June2012.