A Survey on Odia Character Recognition

Pushpalata Pujari, Babita Majhi

Abstract*—Recognition of Odia handwritten and machine characters and numerals is an emerging area of research and finds extensive applications in banks, offices and industries. Very little standard research work has been reported on recognition of handwritten and machine characters and numerals. This paper makes an in depth study on the existing literature on recognition of machine and handwritten Odia characters and numerals. The key steps [44] such as preprocessing, segmentation, feature extraction and classification involved in the recognition process of Odia characters are dealt in details. The well known techniques employed for segmentation, feature extraction and classification tasks of Odia characters are reviewed and their relative strengths and weaknesses are outlined. The paper also discusses the current trends and future research scope in the area of Odia character recognition. It is expected that this paper will be useful to those who will be interested to work in the fields of recognition of Odia characters.

***Index Terms*— Preprocessing, Segmentation, Feature extraction, Classification, Post Processing**

I. INTRODUCTION

Knowledge contained in paper based and handwritten documents are more valuable and beneficial if it is available in digital form. In recent past there are increasing trend to digitize written and paper based documents such as books, newspapers and handwritten materials for the benefit of wider section of the society. It is desirable to preserve these documents in digital forms. The Optical Character Recognition (OCR) is a process [43] by which the printed and scanned documents are converted to ASCII character which is recognized by a computer. The recognition of characters and numeral of a language is a challenging problem since their variations due to different font sizes and different types of variations introduced during writing. The character recognition (CR) can be broadly classified into two groups: offline and on line. In the first case, the document is generated, digitized, stored in memory and then it is processed but in case of online system, the character is processed as soon as it is generated. The factors such as pressure and speed of writing do not influence the offline system, but they effect the online one. Offline and online systems can be applied to handwritten characters (Fig 1(a)) and optical characters (Fig 1(b)) respectively. Accordingly, the recognition task can be classified as OCR or handwritten character recognition [29].

Manuscript Received on February 2015.

Pushpalata Pujari, Computer Science and Information Technology Department, Guru Ghasidas Vishwavidyalaya, Central University, Bilaspur, India.

Babita Majhi, Computer Science and Information Technology Department, Guru Ghasidas Vishwavidyalaya, Central University, Bilaspur, India.

Fig. 1(a) Handwritten Character

Fig. 1(b) Optical Character

In recent past intensive research studies have been undertaken for recognition of characters in various languages, but little reported materials is available in Odia Character Recognition. In this paper a sincere attempt has been made to critically review the existing literature and to provide the latest trend on the recognition of Odia character. An in depth study of Odia character recognition, outlining existing methods and their limitations has been carried out in this paper and future research direction in this field has been suggested. The features of Odia characters are analyzed in Section II. The various steps of character recognition such as image acquisition, preprocessing, segmentation, feature extraction and post processing have been dealt in Section III. Finally, the outcome of the investigation and future research direction are presented in Section IV and V respectively.

II. FEATURES OF ODIA SCRIPT

Around 11th century A. D. [50], [18] the Odia script has been originated from the Brahmi script and has undergone through many transformations. In the year 2013, the Government of India, on the recommendation of Government of Odisha has renamed the state as Odisha and the corresponding language to Odia from Oriya. The Odia is used to write the Odia language which is spoken in Odisha state situated in the eastern part of India. Character and numeral recognitions are required for the development of electronic libraries, multimedia databases and banking systems. The cursive shapes of the Odia letters appear to be influenced by Southern scripts. The cursive shape of the letters may be due to the need to write on palm leaves with a pointed stylus which has a tearing tendency if more straight lines are used. The writing style of Odia script is from left to right. The Odia alphabets are grouped into 12 vowels, 35 consonants and ten numerals. These are called basic or main characters in Odia language. Almost half of these characters contain a straight line on the right side. Some of these characters, mostly vowels, are derived from other basic characters. Sometimes consonants are also combined with consonant to form new character. Special symbols which do not touch the consonant character [18] known as *matras* are added to the consonants. The components of the characters are classified into (a) Main component in form of a vowel or consonant. (b) Vowel Modifier when a vowel following a consonant takes a modified shape and is placed at the left, right (or both) or bottom of the consonant. (c) Consonant modifier when a symbol consists of two or more consonants, the main component and consonant modifier/s or half consonant. In spatial sense, the consonant modifier is employed at the bottom or top of the main component. In practice more than

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

two to four consonant-vowel combinations are available. The resultant character is known compound character or conjuncts. It is observed that the symbol modifiers, vowel modifiers (*matra*) or consonant modifiers have specific positions with respect to the base character [9]. The presently used Odia vowels and consonants with their respective pronunciation are shown in Figs. 2(a) and 2(b) respectively. The commonly used Odia conjunctions with their respective pronunciations are shown in Fig.3. The ten basic numerals with their pronunciation and corresponding English numerals are shown in Fig. 4.

Fig. 2(a) Odia vowels

Fig. 2(b) Odia Consonants

Fig. 3. Some commonly used Odia conjunctions

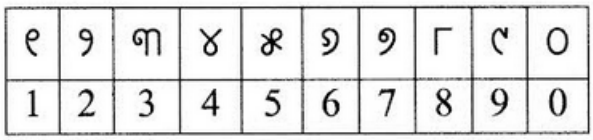

Fig. 4. Odia numerals and their corresponding English numerals

A text in Odia script is mainly divided into three zones: upper zone, middle zone and lower zone. The portion lying between mean line and upper line constitutes the upper zone. The middle zone consists of the area below the mean line and above base line. The portion between base line and lower line comprises of lower zone where some of the modifiers are placed. The imaginary line separating the middle and lower zone is known as the base line. The line which divides the upper zone from middle zone is called the mean line [9], [47]. An illustration of zoning is shown in Fig. 5.

Fig. 5. Identification of different zones and lines of an Odia text line

III. Basic Steps of Odia Character Recognition

Character recognition [34] refers to a technique which enables to transform different documents, such as scanned paper, PDF files or images captured by a digital camera and handwritten documents into editable and searchable form. Review of literature on Odia character recognition reveals that most of the character recognition methods involve few major steps such as image acquisition, preprocessing, segmentation, feature extraction, classification and post processing. Fig.6. depicts a block diagram indicating the major steps involved in character recognition system. A brief outline of the CR is provided in sequel.

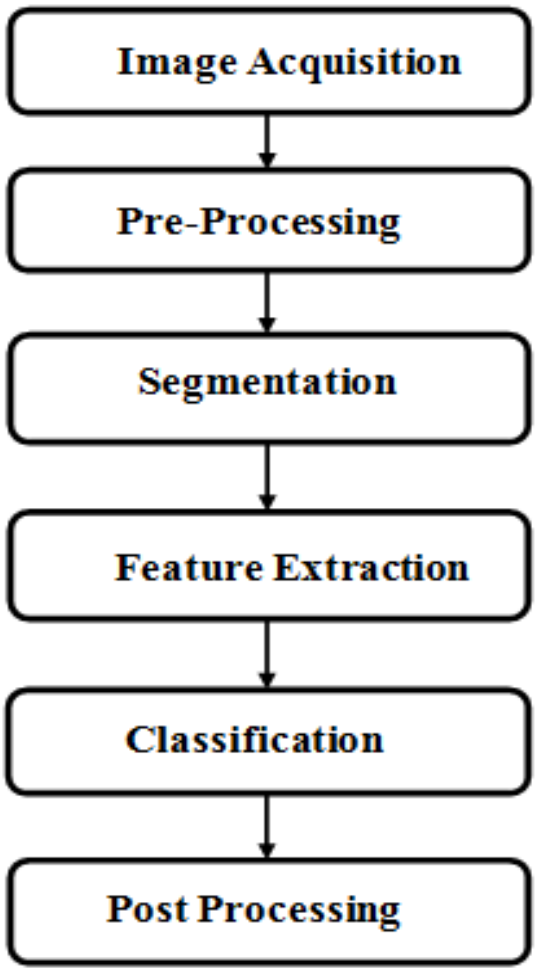

Fig. 6. Basic steps involved in character recognition system

A. Image Acquisition

In the image acquisition or digitization process the images for CR system are acquired by appropriate scanning of handwritten documents, books, and magazines or by capturing photographs of document or by directly writing in computer. The input image is obtained by camera or through some scanner. The images are represented in the formats such as JPEG, BMT, BMP, TIF and TNG. The input acquired may be in gray, color or binary tone [48]. From literature review it is observed that most of the authors have used flat bed scanner with 300 dpi, gray tone for image acquisition. Table I lists various sources of data, number of samples, tone, dpi and formats used by different authors working in the same area and have been reported in the literature.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

Table I. Sources and characteristic of Data

Reference No.	Techniques used	No. of Training Patterns Used	No. of Testing patterns Used	Accuracy Results in percentage
[2]	Feature based tree classifier, run-number based matching approach	Not Reported	Not Reported	96.3 %
[3]	ANN , GA	Fifteen sheets of characters	Five sheets of characters	94 %
[5]	Hopfield NN	1500 characters	290 charcaters	95.4 %
[6]	ANN	1200 Patterns	1100 Patterns	85.30 %
[9]	ANN	1500 Words	1000 Words	97.69 %
[10]	Tesseract OCR	8 data files	More than 1 page	100 %
[11]	AMA	12 Characters	Not Reported	92.42-97.87 %
[14]	BPNN, SVM	15440 samples	4560 samples	i)83.33 %(BPNN),93.4 %(SVM) by using Fourier descriptor feature ii)83.63 %(BPNN),93.57 %(SVM) by using Normalized chain code feature
[15]	Quadrant Mean method	12000 samples	4000 samples	93.20 %
[16]	ANN	75% of the data	15% of the data	91.33-99.6 %
[17]	NN, KNN	38 classes of character each class represebnted by 10 templates	1600 characters	82.33 %(NN),72.27 %(k-NN) with normal character features. 41.88 %(NN), 39.41 %(k-NN) with thinned character features.
[19]	BPNN	150 samples	350 samples	92 %with DTC features 82.70 % With DWT featurres
[22]	BPNN	100 samples	Not Reported	91.24 %
[23]	curvature	15552 samples	3638 samples	94.60 %
[24]	ANN	396 data of each numeral	100 data of each numeral	99.3 % with gradient feature, 95.66 % with curvature feature
[25]	HMM	4970 images	1000 images	90.50 %
[26]	NN and Quadratic	Not Reported	3850 data	94.81 %
[28]	Normalization and thinning free automatic scheme	Not Reported	3550 data	97.74 %

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

B. Preprocessing

The method of extraction of text from the document is called pre-processing or document analysis. The pre-processing consists of a series of operations performed on the scanned input image, which includes background noise reduction, filtering, image restoration etc [38], [42]. It primarily enhances the image making it suitable for segmentation. The choice of preprocessing algorithms employed on the scanned image depend on factors such as document and paper quality, resolution of the scanned image, amount of skew present in the image, format and layout of the image and text, types of script and the type of characters used such as printed or handwritten. The preprocessing steps commonly used are noise reduction, binarization, normalization, skew detection and thinning.

1) Noise reduction

The images of the characters usually get contaminated with additive noise introduced from the scanning devices and/or transmission medium [45]. The noise degrades the quality of images which poses problem in the subsequent steps of character recognition system. The presence of noise introduces disconnected line segment, large gaps between the lines etc [38]. The device or the writing instrument sometime introduces noise in the form of disconnected line segments, bumps and gaps in lines, filled loops, etc. The distortions in form of local variations, rounding of corners, dilation, and erosion are also matters of concern. It is required to get rid or reduce these effects before the CR task is carried out. Three common types [8] of noise occur in handwriting are background noise, shadow noise and salt and pepper noise. Smoothing operation is carried out to eliminate the artifacts present during capturing of noise. Two main methods used for this purpose are filtering through masking and morphological operation such as erosion, dilation. The noise removal methods eliminate unwanted bit-pattern which do not contribute to the output. Common filters like the mean filter, min-max filter, Gaussian filter have been applied to remove noise from the documents . A logical smoothing approach is suggested in [2] for filtering out digitized image with protrusions, dents in the characters and isolated black pixels over the background.

2) Binarization

The process of conversion of a gray scaled image into binary image is known as binarization which is achieved by thresholding. In this operation the gray scale or color images are changed to binary images by choosing a suitable thresholding value [38], [41], [42]. In order to achieve less storage and to increase rate of processing the gray scale or color images is changed to binary images. The thresholding operation extracts the foreground from the background. The histogram of grayscale values of a typical document image represents one peak corresponding to the foreground and another peak for the white background. Therefore the threshold value is chosen in the valley between the two peaks. The thresholding may be of three types: global, local and hybrid. In global thresholding one threshold value for the entire document image is chosen from the estimated value of background level from the intensity histogram of the image. In local thresholding different threshold values are employed for different regions of the image [8]. Hybrid thresholding technique attempts to combine the advantages of global and local thresholding. It makes better adaptability of various types of noise at different areas of the same image with less computation and time [2]. The adaptive threshold method is used by Mishra et.al [1] for converting grayscale image into binary image. Chaudhuri et.al [2] have chosen a histogram-based thresholding technique to transform images. For the white and black regions of the document the histogram shows two distinct peaks. The midpoint between the two histogram peaks is selected as the threshold value. The two-tone image is converted into two labels where 1 and 0 corresponds to object and background respectively. Thresholding technique is also used in [3] to convert the grayscale image to binary image. A two-stage approach is suggested in [9] to convert into two-tone (0 and 1) image. The first stage comprises of pre binarization which uses a local window based algorithm to obtain different regions of interest. The run length smoothing algorithm (RLSA) is then used on the gray scale image. Subsequently histogram based global binarization is employed to get the binary image. In another communication [15] the gray levels are scaled to fall within the range 0 to 1 without performing any skeletonization. For a eight-bit representation a threshold value is suitably chosen [22] and any value above it is chosen as 1 otherwise it is represented as 0.

3) Normalization

Normalization is carried out during pre processing stage to remove all types of variations present in the image and to obtain a standard size data [39][46]. The scaling, translation, and rotation etc. constitutes various steps of normalization. It is employed to avoid scaling and rotational effects. Document images are invariably different in sizes and the algorithms are applied on a fixed size image matrix. The documents are usually normalized with respect to width, height or both. For comparison of the performance normalized documents should be used [20]. The size of segmented digits/numerals varies typically around 200 to 256 pixels .A linear transformation has been proposed [15] to preserve the aspect ratio of the character. In some cases [16] normalization of the character has been done to achieve zero mean and unit standard deviation. Such standardization of the input image makes it independent from the size. In [19] and [24] the Odia numerals have been normalized to sizes 256 X 256 and 64 X 64 respectively. For obtaining a gray scale image a mean filter of size 3 X 3 has been repeatedly applied. The gray scale image is further normalized to achieve zero mean and unity maximum gray scale value.

4) Skew detection

The skew or tilt is the deviation of the baseline text from horizontal direction [2], [41]. When a document passes through the scanner mechanically or by an operator, a few degrees of tilt is observed. While saving the scanned document it may have some tilt or in other words the image may have under gone some rotation. Detection and correction of these tilts are important preprocessing steps in document analysis. The skew may be corrected in two steps: (i) estimation of tilt angle and (ii) rotation of the image by the same amount in the opposite direction. A transform based approach has been introduced [2] for estimating the tilt of Odia documents. The accuracy of skew detection affects

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

segmentation and classification steps of character recognition. Skew detection is required to align the text or document image to coordinate axes [38]. The image is rotated based on the detected skew angle. In this case the estimation method is not affected by the font style and variation in size. This approach is not limited to the range of the tilt. The algorithm proposed in [35] is employed in [9] to de-skew the documents .The projection profile and the Winger-Ville distribution (WVD) [17] techniques have been used for the skew angle estimation. A skew detection algorithm is proposed in [37, 40].

5) Thinning

It is a process which removes the undesired pixels and transforms the image pattern one pixel thick by retaining the connectedness of the object and its end points. When no additional pixel can be further removed from the image the thinning operation becomes complete [36],[41]. It is a morphological operation which removes selected foreground pixels from the binary images. This data reduction process erodes an object until it is one pixel wide and produces a skeleton of the object making it simpler to recognize [38]. Edge detection is a process of locating points in a digital image at which sharp change in the brightness occurs [1]. Bhagirathi et.al [11] have employed connected component analysis for thinning. It helps to remove the thickness effect of the pen used for writing.

C. Segmentation

It is an important stage [45] as it enables separation of words, lines, or characters directly and hence it affects the recognition rate of the script. Two types of segmentation [39] are used: external and internal. In external segmentation various writing units, such as paragraphs, sentences, or words are isolated where as the internal segmentation enables isolation of letters, especially in cursively written words. The image is first subdivided into many parts for easy reading . To carry out this task the image is divided in three ways such as line wise segmentation, word wise segmentation, and finally character wise segmentation [46]. In case of line segmentation, the image is divided into the line which enables reading of the image limited to lines. For word wise segmentation these lines are further divided into words which allows understanding of image restricted to the words in lines and by such operation small blocks called words get separated. For achieving character wise segmentation, the image is further divided and the algorithm separates the document image into more small blocks called characters [36]. Several methods of segmentation have been proposed in the literature. Chaudhuri et.al [2] have proposed to count the number of black pixels in each row of the lines of a text box by finding the valleys of the projection profile. The process of boundary detection and dilation is employed in [3] to segment word from a complete page of handwritten text and the characters are extracted from the words. The concept of water overflow from a reservoir has been suggested to segment the touching characters of Odia handwritten text into individual characters [7]. In this scheme the text image is first segmented in to lines, and the lines are then segmented into individual words. Then the isolated and connected (touching) characters in a word are detected for character segmentation. Characters of the touched word are usually segmented using structural, topological and water reservoir concept. They have used a piece-wise projection method to take care of unconstrained handwritten documents. The density of black pixels has been computed with the candidate length of the line to take care of wrongly segmented lines. First, they have detected isolated and connected (touching) characters within each word to segment characters from words. Subsequently the connected components are segmented into individual characters [8]. In this work the handwritten text is first divided into lines. The spacing between the words is used for word segmentation. By taking the vertical connecting pixel (VCP) of an input text line the spacing between the words is obtained. For character segmentation the spacing between the characters in a word is used. In [9], a piecewise projection method is used for segmenting a document into lines and then lines into words. Then the vertical histogram of the line for word segmentation has been computed. In general, the distance between two consecutive words of a line is more than the distance between two consecutive characters in a word. Considering the vertical histogram of the line and using distance criteria the words have been segmented from lines. In [15], the segmentation scheme has been implemented in two steps. A two-stage artificial network based classifier is designed for a coarse classification between textual document and PIN-code numeral. Subsequently a fine classification scheme is implemented to separate each numeral of the PIN-code. Employing linear transformation the characters are then normalized. The projection of the image into the vertical axis and horizontal projection for line segmentation are proposed in [18] and [22]. In these works the words and characters in a line have been separated using the projection of the segmented line on to the horizontal and vertical axes. To separate the *matras* situated either above or below the character, a connected component analysis have been employed. In another study [20] the lines have been segmented based on gray scaled image. The segmented text lines have been used as the input for the word segmentation method which produces segmented words. Both foreground and background information is used in this work. In another communication [26], accurate line segmentation has been achieved for Odia text printed documents. It produces the output as text line segment of Odia file. Meher.S, .D [49] segmented text into lines and then each line is segmented into individual words and then each word is segmented into individual characters or basic symbols. The spacing between the words is used for word segmentation by taking the Vertical Connecting Pixel (VCP) of an input text line. Spacing between the characters in a word is used for character segmentation. Segmentation results reported in the literature with different number of samples in each category is shown in Table II.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

Table II. Comparison of accuracy obtained on line, word and character segmentation

Reference No	Number of Lines	Accuracy	Number of Words	Accuracy	Isolated and Connected Characters	Accuracy	Touching Component	Accuracy
[2]	Not Reported	97.5 %	Not Reported	97.7 %	97.2 %	97.2 %	Not Reported	Not Reported
[7]	1627 lines	97 %	3700 Words	98.2 %	3200 characters	96.3 %	1428(Two Characters) 311(Three Characters) 71(More than three characters)	96.7%(Two Characters) 95.1%(Three Characters) 93.3%(more than three Characters)
[20]	5088 lines	99.3 %	57224 words	86.5 %	Not Reported	Not Reported	Not Reported	Not Reported

D. Feature Extraction

Each character or numeral contains some special and distinct characteristics to uniquely represent it. To find a set of parameters that uniquely defines the character is called feature extraction. The feature extraction technique should be such that the features of characters should enable clear discrimination of one character from others. To distinguish a class from other class a set of features is extracted for each class. The types of feature may be of statistical, syntactical/structural or hybrid in nature. Statistical features are obtained by computing the statistical and geometrical moments. The structural features are represented by strokes, holes, end points, loops or cross-over points. The hybrid features constitute suitable combination of statistical and structural features [36]. The two important sub-stages of recognition are feature extraction and classification. In the feature extraction stage a text segment is analyzed and a set of features are computed to uniquely identify the text segment. These features are then used as input to the character classifier. In [1] different extracted geometric such as height, width, number of pixels in columns and rows and textual features like histogram and centroid have been used to obtain a valid epoch or score. Topological and stroke-based features as well as features obtained using the concepts of water overflow have been employed for character recognition in [2]. The features are chosen by considering (i) robustness, accuracy and simplicity of detection, (ii) speed of computation, (iii) independence of size and fonts and (iv) need of classifier design. In [3] the feature vector has been generated based on the standard deviation, average angle ad zone based average distance from zone centroid and image centroid. The feature detection methods are simple and robust, and do not involve preprocessing steps like thinning and pruning. Hu's seven moments and Zernike moments have been successfully used in [4] to extract the features of Odia characters. The features comprising of shape, size and position of a digital curve with respect to the numeral image has been used in [5]. LU decomposition of matrix [6] have been employed to extract the feature vectors from the character image. In [9] the features using (i) fractal based feature, (ii) water reservoir based feature, (iii) presence of small component, (iv) topology have been extracted from the characters for recognition. A suitable combination of the global and local (zone based) features have been used in [11]. The global features constitute numbers of loops, end points, horizontal strokes, vertical strokes, angular strokes and the aspect ratio. The local features are the number of cross-point with three and four connections. They measure the center of gravity (CoG) of the cross and the end points. A recent paper [14] has dealt in offline recognition of isolated Odia handwritten numerals using Fourier descriptors and normalized chain code as features. In this paper the features have been extracted from the shape of the binary image of the numerals. Another novel feature extraction reported [15] is based on splitting a numeral image into four quadrants and then taking the mean of the gray values of pixels of each quadrant. In a recent communication [16] features have been obtained from the directional information of the characters. For computing the features, the bounding box of a numeral is segmented into blocks and gradient in each direction is computed for each of the blocks. A Gaussian filter is then used for down sampling the blocks. The steps followed for feature extraction are binarization, normalization, and use of Robert's cross operator and Gaussian filter. Pati et.al [18] have normalized the test character to a given size based on the aspect ratio (character width / character height). The normalized character is then divided into number of rectangular sectors. Second order geometric moments are extracted from each of these sectors to represent the feature vector. Spatial features and neural networks have been proposed in [19]. After classifying into two groups, all the characters are resized into 20×14 pixels. Each resized character contains 280 pixels which serve as features for training the neural networks. Gradient and curvature features have been used in [23] for recognition purpose. For feature extraction at first the image is normalized and this normalized image is then segmented into 49 x 49 blocks. A best trade-off between accuracy and complexity is achieved by suitably fixing the size of the blocks. To achieve strength and direction of gradient they have applied Roberts filter on the image. By using bi-quadratic interpolation method curvature features have been computed. Feature extraction process consisting of gradient calculation, curvature computation, feature vector generation and dimension reduction of the feature vector are reported in [24]. The shapes of the strokes have been analyzed for extraction of features in [25]. For example, from each stroke in *E* and *S*, eight scalar features are extracted. These features indicate the shape, size and position of a digital curve

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

with respect to the numeral image. Roy et.al [26] have used histograms of direction chain code of the contour points of the numerals as features for recognition task. In [28] water reservoir concept as well as topological and structural features of the numerals have been considered. Reservoir based features such as number of reservoirs, their sizes, heights and positions, water flow direction and topological features like number of loops, center of gravity, positions of loops, the ratio of reservoir/loop height to the numeral height, profile based features, feature based on jump discontinuity have been suggested in the recognition scheme. A standard deviation and zone centroid average distance based feature matrix have been considered in [31]. Mitra et.al [33] have extracted directional features by directional decomposition of character image and using fixed scheme. Sanghamitra Mohanty [50] used Feature weighting on basis of longest- run features with respect to wrapper-based feature weighting algorithm for K K - Nearest Neighborhood.

E. Classification

The classification stage represents the decision making part of a recognition system and it employs the features extracted in the previous stage as inputs to the classifiers. The classifiers compare the input features with the stored features to assign a class for the input. The classification can be broadly divided into methods based on statistics, artificial neural networks (ANNs), kernel, and multiple classifier combination. Character classifier can be grouped as Baye's scheme, nearest neighbor classifier, radial basis function, support vector machine, artificial neural network etc. The character recognition task is based on four approaches as template matching; statistical techniques; structural techniques and neural network. The overall performance of the recognition system depends mainly on the type of the classifier used. The multi layer artificial neural network has been proposed [1] for efficient recognition. In [2] the Odia characters sets have been classified into four groups according to similarity of their shapes and features. Recognition of the basic and compound characters have been carried out in two stages. By using a feature based tree classifier the characters are first grouped into small subsets. In the second stage, using a sophisticated run-number based matching approach; the characters in each group are recognized. In [3], the feed forward BPNN algorithm and the genetic algorithm (GA) have been proposed to perform the optimum feature extraction and recognition. The Odia character sets have been classified into four groups according to similarity of their shapes and features. Five ANNs having different parameters have been used and their outputs are used to choose the best solution using the GA based optimization. A Zernike moment based approach has been applied in [4] to recognize Odia characters. Sarangi et.al [5] have employed Hopfield neural network (HNN) model to recognize the handwritten Odia characters. A total of 290 characters have been used to test the recognition ability of HNN. The hand written numerals and handwritten characters in Odia language have been recognized using multilayered perceptron network in [6] and [9] respectively. A substructure based method has been proposed in [8] for Odia character recognition. To recognize printed documents of Odia language Tesseract OCR engine is used in [10]. The ant-miner algorithm (AMA) has been introduced in [11] for offline OCR of hand written Odia scripts. The AMA is a rule-based approach used for training proposes. The authors have defined three types of block as per the writing styles of the scripts. The performance of AMA is then tested with four characters from each block. Finally, a character recognition tool has been developed for observation and validation. An automatic image processing approach using morphological technique for extracting individual characters for training set as well as an artificial neural network for classification is proposed in [12]. In another communication [13], the combination of Naïve Bayes classifier with LU factorization on a smaller size of data set has been used. The support vector machines (SVM) is used in [11] for classification and recognition purpose. Mahato et.al [15] have employed two stage artificial neural network based general classifiers for the recognition of PIN-code digits written in Odia. In this paper they have used a novel technique known as quadrant- mean technique to identify the numerals of PIN code written in Odia script. The multilayered perceptron neural network is used for the recognition of numeral characters. In [16], ANN classifier has been used for recognition. Various other related works reported in the literature are the k-nearest neighbors (KNN) method [18] and BPNN methods [19] of recognition. In [23], curvature features have been used for the recognition purpose. A low complexity single layer neural network is proposed for classification of Odia numerals in [24]. The gradient and curvature features of the numerals have been taken as the inputs to the functional link artificial neural network (FLANN) classifier. In [25], a novel hidden Markov model (HMM) has been suggested for recognition of handwritten Odia numerals. To classify an unknown numeral image, its class conditional probability for each HMM is computed and is used for recognition. The neural network (NN) and quadratic classifiers separately have been introduced [26] for recognition purpose. The KNN and Khonen feature map have been suggested for Odia character recognition in [27]. In a recent paper [28], normalization and thinning free automatic scheme for unconstrained off-line Odia isolated handwritten numeral recognition has been proposed. In [31], a feed forward BPNN algorithm in two stage is employed to perform the optimum feature extraction and recognition. The system employes the ANN in two stages, having different configurations, the first stage classifies the characters into similar groups and in the second stage individual characters are recognized. Nigam et.al [32] have proposed a new character recognition method for Odia script based on curvelet transformation based coefficients. Multi-class SVM classifiers have been proposed in [33]. They have used a simple but novel directional decomposition technique for recognition of printed Odia characters. Odia characters are circular in nature but most of the distingishing information occurs in non-circlar portions. The relative position and orientation of linear strokes are exploited to distinguish individual characters. Table III shows the performance of different techniques in terms of number of training and testing patterns used and classification accuracy repoted in the literature.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

Table III. Performance comparision of different classification techniques

Ref. No.	Techniques used	No. of Training Patterns Used	No. of Testing patterns Used	Accuracy Results in percentage
[2]	Feature based tree classifier, run-number based matching approach	Not Reported	Not Reported	96.3 %
[3]	ANN , GA	Fifteen sheets of characters	Five sheets of characters	94 %
[5]	Hopfield NN	1500 characters	290 charcaters	95.4 %
[6]	ANN	2200 Patterns	1100 Patterns	85.30 %
[9]	ANN	1500 Words	1000 Words	97.69 %
[10]	Tesseract OCR	8 data files	More than 1 page	100 %
[11]	AMA	12 Characters	Not Reported	92.42-97.87 %
[14]	BPNN, SVM	15440 samples	4560 samples	i)83.33 %(BPNN),93.4 %(SVM) by using Fourier descriptor feature ii)83.63 %(BPNN),93.57 %(SVM) by using Normalized chain code feature
[15]	Quadrant Mean method	12000 samples	4000 samples	93.20 %
[16]	ANN	75% of the data	15% of the data	91.33-99.6 %
[18]	NN, KNN	38 classes of character, each class represebnted by 10 templates	1600 characters	82.33 %(NN),72.27 % (k-NN) with normal character features. 41.88 % (NN), 39.41 % (k-NN) with thinned character features.
[19]	BPNN	150 samples	350 samples	92 % with DTC features 82.70 % With DWT featurres
[22]	BPNN	100 samples	Not Reported	91.24 %
[23]	Quadratic Classifier	Not Reported	18190 samples	94.60 %
[24]	ANN	396 data of each numeral	100 data of each numeral	99.3 % with gradient feature, 95.66 % with curvature feature
[25]	HMM	4970 images	1000 images	90.50 %
[26]	NN and Quadratic	Not Reported	3850 data	94.81 %
[28]	Normalization and thinning free automatic scheme	Not Reported	3550 data	97.74 %

One important observation from the classification accuracy reported in literature and listed in Table 3 is that the results obtained are not using any standard data bases handwritten machine characters and numerals.

F. Post processing

The goal of post processing phase refers to detect and correct linguistic misspellings in the OCR output text after the input image has been completely processed [29], [30]. Post processing steps are used to improve the accuracy of OCR character recognition system. It is very difficult to process the data which contains spelling mistakes. The accuracy of OCR can be increased if the output is constrained by a lexicon – a list of words that are allowed to occur in a document. Higher level analysis such as syntax and semantic analysis can be applied to check the context of the recognized characters. Post processing phase can be broadly divided into three groups: manual error correction, dictionary-based error correction, and context-based error correction. Manual error correction requires a continuous manual human intervention. In this case the correction is made manually. In dictionary based error correction, a lexicon or a lookup dictionary is employed to spell check the OCR recognized words and correction is made if they are misspelled. Context-based error correction

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

techniques perform error detection and correction based on the grammatical error and semantic context. From literature survey it is learnt that a little work has been done on post processing.

IV. CONCLUSION

This paper presents a thorough and up-to-date review of Odia character recognition. The research work carried out during the last decade in the field of Odia character recognition has been surveyed. Different approaches employed for each step of Odia Character recognition are also outlined. Each of these methods has its own advantages and limitations. It is observed that the broad steps in a typical character recognition system comprises of image enhancements, noise removal, skew detection, binarization, normalization, segmentation, feature extraction and classification. Each step contributes to the overall accuracy of the system. Line and word segmentation are initial requirements for the OCR system. The text lines and the words in a Odia document are segmented because the recognition process begins. Character recognition task becomes simpler if the zones are distinguished because the lower zone contains modifiers and the *halant* mark, whereas in the upper zone the modifiers and portions of some basic characters lie. It is important to extract distinct features before recognition task is carried out..Literature review reveals that mostly neural network has been chosen as classifiers. Combinations of artificial neural networks and genetic algorithms have been reported to provide some satisfactory results. Nearest neighbor classifiers require less storage space and computation time than that of the SVMs. It is observed that the curvelet based method yields highest accuracy and efficiency than other traditional methods of feature extraction. The results of the quadratic classifier provide improved recognition performance compared to that of NN classifier. BPNN based system for recognition of handwritten Odia characters provides satisfactory performance compared to other methods. By reducing the number of input vectors to the neural network, the computation time can be decreased. It is also observed from the literature that work on post processing phase, which is crucial to discriminate similar structured characters, is very few. Every step of Odia character recognition is important as each of them contributes to overall performance of the system.

V. FUTURE RESEARCH

Recognition of character is still a challenging problem since there is a variation in same character due to different font size, different types of noises and involvement of different persons. During the last decades, intensive research studies have been made for recognition of handwritten characters and numerals in various Indian and foreign languages, but a few work has been reported on Odia character recognition. The field of character recognition in Odia language still needs an in depth study.

- Research work is needed to develop Bi-lingual OCR using Odia as one.
- Performance of various nonlinear classifiers such as the ANN, BPANN, KNN, HNN, CNN, SVM and HMM need to be evaluated for each case to choose the best method for Odia hand written numerals and characters
- In most of the cases the accuracy of recognition is obtained by taking a limited set of samples. The true accuracy rate can be assessed by taking different set of samples for training and testing purposes.
- The accuracy of recognition reported on Odia numerals and characters may not hold good for large and unique database. Hence there is a need of evaluation of unified study.
- A complete OCR system is required to convert one page of text to its ASCII format.
- In most of the cases the accuracy of recognition is severely affected due to presence of similar shaped Odia characters .Recognition of such characters requires sincere effort.
- Even though there are number of publications in conferences on Odia character and numeral recognition, very few material on standard journal are available.
- To identify the touching characters uniquely is a challenging task. This issue needs further study to achieve effective solution.
- The Odia OCR should be developed to accurately recognize characters in diverse fonts and sizes.
- Because of non-uniform writings, some words are printed closer and are not separated equally, some time it is difficult to correctly separate touching characters .This issue also requires attention and further study.
- The misclassification of characters and numerals obtained from the classifiers could be due to the local minima problem associated with the learning algorithm. Evolutionary computing algorithms such as the particle swarm optimization (PSO) and bacterial foraging optimization (BFO) can be employed for better training of the classifier.
- In essence, As there are no standard databases presently available for Odia handwritten characters more research effort is necessary for creation of good standard data base, enhancing appropriate features by adopting proper feature extraction methods and finally developing acceptable, robust hybrid classifiers. The resulting Odia OCRs will then be more useful for many real life applications.

REFERENCES

[1] Soumya Mishra, Debashish Nanda , Sanghamitra Mohanty , Oriya Character Recognition using Neural Networks , Special Issue of IJCCT Vol. 2,3,4, 2010 for International Conference (ICCT-2010), pp. 88-92.
[2] B. B. Chaudhuri, U. Pal and M. Mitra, Automatic recognition of printed Oriya script, Sadhana (27) (Part 1) (February 2002), 23–34.
[3] Debananda Padhi, Novel Hybrid approach for Odia Handwritten Character Recognition System ,International Journal of Advanced Research in Computer Science and Software Engineering , 2 (5) (May 2012) 150-157.
[4] Jyotsnarani Tripathy, Reconstruction of Oriya Alphabets Using Zernike Moments, International Journal of Computer Applications, 8(8) (October 2010) 26-32
[5] Pradeepta K. Sarangi, Ashok K. Sahoo, P. Ahmed, Recognition of Isolated Handwritten Oriya Numerals using Hopfield Neural Network, International Journal of Computer Applications , 40(8) (February 2012) 37-42.
[6] Pradeepta K. Sarangi, P. Ahmed, Recognition of Handwritten Odia Numerals Using Artificial Intelligence Techniques, The International Journal of Computer Science & Applications (TIJCSA), 2(2) (April 2013) 41-48.
[7] N.Tripathy and U. Pal, Handwriting segmentation of unconstrained Oriya text, Sadhana, 31(6)(December 2006), 755–769.
[8] Debasish Basa and Sukadev Meher, Handwritten Odia Character Recognition, National conference on Recent Advances in Microwave

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

tubes , Devices and Communication, System ,JNIT , Jaipur, March4-5 2011.

[9] K. Roy and U. Pal, Word-wise Hand-written Script Separation for Indian Postal automation, In : Proceedings of 10th International Workshop on Frontiers in Handwriting Recognition , (2006) 521-526.

[10] Mamata Nayak, Ajit Kumar Nayak, Odia Characters Recognition by Training Tesseract OCR Engine, International Conference in Distributed Computing & Internet Technology (ICDCIT-2014), International Journal of Computer Applications (2014) 25-30.

[11] Bhagirath Kumar, Niraj Kumar, Charulata Palai, Pradeep Kumar Jena, Subhagata Chattopadhya, Optical Character Recognition using Ant Miner Algorithm: A Case Study on Oriya Character Recognition, International Journal of Computer Applications , 61(3) (2013)17-22.

[12] Debananda Padhi, Debabrata Senapati, Sasmita Rout, Morphological Approach for Extracting Single Character from Odia Handwritten Text: A survey, International Journal of Emerging Trends in Engineering and Development (IJETED) , 2(2)(2012) 138-146.

[13] Pradeepta K. Sarangi, P. Ahmed and Kiran K. Ravulakollu, Naïve Bayes Classifier with LU Factorization for Recognition of Handwritten Odia Numerals, Indian Journal of Science and Technology, 7(1) (January 2014)35-38.

[14] Rasmi Ranjan Das, Swati Sucharita Das, Shom Prasad Das, Support Vector Machines for Odiya Handwritten Numeral Recognition, International Journal of Advanced Research in Computer Science, 4(9) (2013),139-143.

[15] Manoj Kumar Mahto, Archana Kumari and S. C. Panigrahi, A System for Oriya Handwritten Numeral Recognition for Indian Postal Automation, International Journal of Applied Science & Technology Research Excellence 1(1)(Nov-Dec 2011) 17-23

[16] Priyaranjan Behera, Odia Offline Character Recognition,Thesis,2012, http://ethesis.nitrkl.ac.in/3823/1/Thesis__Odia_Offline_Character_Recognition__108CS021.pdf , Access Date : 12/09/14

[17] Avijeeta Mohanty , Debananda Padhi, Soumya Mishra, A Novel WVD Approach for Estimating and Correcting Skew angle of Odia Handwritten Document Image, International Journal of Advanced Research in Computer Science and Software Engineering, 2(3)(2012)175-181

[18] Peeta Basa Pati , A.G.Ramakrishnan, U.K.Aravinda Rao, Machine Recognition of Printed Oriya Characters,In: Proceedings of III International Conference on Information Technology ICIT 2000, Bhubaneshwar, Decemeber 21-23, 2000, pp. 227-232.

[19] T.K.Mishra, B.Majhi, S.Panda, A comparative analysis of image transformations for handwritten Odia numeral recognition, In : proceedings of IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI), Mysore,22-25 Aug, 2013, pp. 790-793.

[20] D. Senapati, S. Rout and M. Nayak, A Novel Approach to Text Line and word Segmentation on Odia Printed Documents , In : Proceedings of IEEE Third International Conference on Computing Communication and Networking Technologies 2012, 6th - 28th July 2012, pp.1- 6.

[21] Sukhpreet Singh, Optical Character Recognition Techniques: A Survey, Journal of Emerging Trends in Computing and Information Sciences, 4(6)(June 2013)545-550.

[22] S.D.Meher and D. Basa, An Intelligent Scanner with Handwritten Odia Character Recognition Capability ,In: proceedings of IEEE Fifth International Sensing Technology(ICST), Palmerston North ,Nov 28 2011-Dec.1 2011, pp 53-59.

[23] U.Pal, T. Wakabayashi, F.Kimura , A System for Off-Line Oriya Handwritten Character Recognition Using Curvature Feature, In: Proceedings of 10th International Conference on Information Technology(ICIT) 10th ,Orissa, 17-20 Dec.2007, pp: 227-229.

[24] B. Majhi, J. Satpathy , M.Rout, Efficient Recognition of Odia Numerals using Low Complexity Neural Classifier, 2011, IEEE International Conference on Energy, Automation and Signal (ICEAS), Bhubaneswar,30-Dec,pp.1-4

[25] T. K. Bhowmik, S. K. Parui, U. Bhattacharya and B. Shaw, An HMM based Recognition Scheme for Handwritten Oriya Numerals,In: Proceedings of the 9th International Conference on Information Technology(ICIT), Bhubaneswar,India, S. P. Mohanty & A. Sahoo (Eds), IEEE Computer Society Press, December 18-21, 2006, pp. 105-110.

[26] K.Roy, T.Pal, U.Pal, F.Kimura, Oriya handwritten numeral recognition system,In: Proceedings of IEEE Eigth International Confernece on Document Analysis and Recognition ,29 Aug-1 Sept,2005, pp.770-774.

[27] S. Mohanty, Pattern Recognition in Alphabets of Oriya Language using Kohonen Neural Network, International Journal on Pattern Recognition and Artificial Intelligence , 12(07),(November 1998) 1007-1015.

[28] N. Tripathy, M. Panda, U. Pal, System for Oriya handwritten numeral recognition, In: Proceedings of Document Recognition and Retrieval XI, San Jose, California; December 15, 2003; pp. 174-181.

[29] Mansi Shah and Gordhan B Jethava , A Literature Review on Hand Written Character Recognition ,Indian Streams Research Journal ,3(2)(2013) 1-19.

[30] Youssef Bassil and Mohammad Alwani, OCR Post-Processing Error Correction Algorithm U sing Google's Online Spelling Suggestion, Journal of Emerging Trends in Computing and Information Sciences, 3(January 2012),90-99 .

[31] Debanandan Padhi and Debabrata Senapati, Zone Centroid Distance and standard Deviation Based Feature Matrix for Odia Handwritten Character Recognition, In: Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA), Advances in Intelligent Systems and Computing ,Springer 199(2013) 649-658.

[32] Swati Nigam and Ashish Khare, Multifont Oriya Character Recognition using Curvelet Transform, Information systems for Indian languages, Communication in computer and information science, 139(2011)150-156.

[33] Chandana Mitra, Arun K. Pujari, Directional Decomposition for Odia Character Recognition, Mining Intelligence and Knowledge Exploration , Lecture Notes in Computer Science, Springer, 8284 (2013) 270-278.

[34] C. Bihari, Babita Majhi and G. Panda, A critical review on offline handwritten Odia character recognition techniques, In: Proceedings of International Conference on Emerging Trends in soft Computing and ICT, GG Central University, Bilaspur, 16-17, March 2011, pp.86-89.

[35] Z. Shi and V. Govindaraju, Skew Detection for Complex Document Images/using Fuzzy Runlength, In: Proceeding of 7th ICDAR, 2003, pp. 715-719.

[36] Mamta Maloo, K.V. Kale, Gujarati Script Recognition: A Review, International Journal of Computer Science (IJCSI) ,8(4)(No 1)(July 2011) 480-489.

[37] K. Mahata, Optical Character Recognition for Printed Tamil Script, Master's Thesis, Department of Electrical Communication and Engineering, Indian Institute of Science Bangalore, 2000.

[38] Gaurav Kumar, Pradeep Kumar Bhatia and Indu, Analytical Review of Preprocessing Techniques for Offline Handwritten Character Recognition, International Journal of Advances in Engineering Sciences ,3 (3)(July, 2013)14-22.

[39] Nafiz Arica and Fatos T. Yarman-Vural, An Overview of Character Recognition Focused on Off-Line Handwriting, IEEE Transactions on Systems, man and Cybernetics-Part C: Applicaions and Reviews, 31(NO. 2) (2001) 216-233.

[40] K. Mahata and M.Rama Krishnan, Precision Skew Detection through Principal Axis, In proceedings of International Conference on Multimedia on Processing and Processing, IIT Chennai, Aug 13-15, 2000,pp.186-188.

[41] Iping Supriana*, Albadr Nasution, Arabic Character Recognition System Development, The 4th International Conference on Electrical Engineering and Informatics (ICEEI 2013, Procedia Technology 11 (2 0 1 3) 334 – 34.

[42] Amit Choudhary, Rahul Rish, Savita Ahlawat, "Off-Line Handwritten Character Recognition using Features Extracted from Binarization Technique ",2013 AASRI Conference on Intelligent Systems and Contr, AASRI Procedia 4 (2013) 306– 312.

[43] Hacene Belhadef, Mohamed Khireddine Kholladi, Aicha Eutamene, Ontology of graphemes for Latin character recognition, 2011 International Conference on Advances in Engineerin, a Engineering 24 (2011),579-584.

[44] Anju K Sadasivan, T.Senthilkumar, Automatic Character Recognition in Complex Image, International Conference on Communication Technology and System Design 2011, Procedia Engineering 30 (2012) 218 –225

[45] Amit Choudhar, Rahul Rishi, Savita Ahlawa, A New Character Segmentation Approach for Off-Line Cursive Handwritten Words , Information Technology and Quantitative Management (ITQM2013, Procedia Computer Science 17 (2013) 88 –95

[46] N. Shanthi Æ K. Duraiswamy, A novel SVM-based handwritten Tamil character recognition System, Pattern Anal Applic (2010) 13:173–180, DOI 10.1007/s10044-009-0147-0

[47] Subhadip Basu,, Nibaran Das, Ram Sarkar, Mahantapas Kundu, Mita Nasipuri, Dipak Kumar Basu, A hierarchical approach to recognition of handwritten *Bangla* characters, Pattern Recognition 42 (2009) 1467 -1484

[48] Vijay Laxmi Sahu, Babita Kubde, Techniques using Neural Network: A Review, International Journal of Science and Research (IJSR), Volume 2 Issue 1, January 2013, pp:87-94, India Online ISSN: 2319-7064

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

[49] Meher.S, .D, An intelligent scanner with handwritten odia character recognition capability, Sensing Technology(ICST), 2011 Fifth International Conference, Palmerston North on Nov 28 2011-Dec.1 2011, pp 53-59, ISSN: 2156-8065, Print ISBN:978-1-4577-0168-9, Digital Object Identifier: 10.1109/ICSensT.2011.6137038

[50] Sanghamitra Mohanty, Himadri Nandini Das Bebartta, Performance Comparison of SVM and K-NN for Oriya Character Recognition, (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Image Processing and Analysis,pp:112-116

Babita Majhi, is presently working as an Assistant Professor in the department of CSIT, GGV, Central University, Bilaspur, India. She did her Ph.D. in 2009 from National Institute of Technology Rourkela and Post Doctoral research work at University of Sheffield, UK (Dec.2011-Dec. 2012) under Boyscast Fellowship of DST, Govt. of India. She has guided 02 Ph.D. and 08 M.Tech theses in the field of adaptive signal processing, computational finance and Soft-computing and has published 100 research papers in various referred International journals and conferences. Her research interests are Adaptive Signal Processing, Soft Computing, Evolutionary Computing, Computational Finance, Distributed Signal Processing and Data Mining.

Pushpalata Pujari, is an assistant professor in CSIT department, GGV, Central University, Bilaspur, and Chhattisgarh, India. She received her M.C.A Degree from Berhampur University, Berhampur, Odisha, India in 1998. She is currently pursuing her Ph.D in the department of Computer Science and Information Technology, Guru Ghasidas Vishwavidyalaya, Central University, Bilaspur, Chhattisgarh, India. Her areas of interest include Character Recognition, Soft Computing and Data Mining.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.