Gujarati Script Recognition: A Review

Mamta Maloo[1], Dr. K.V. Kale[2]

**[1] PG Dept. Of Computer Science and Technology, SGB Amravati University,
Amravati (M.S.), India**

**[2] Department of Computer Science and Information Technology, Dr. BAM University,
Aurangabad (M.S.) ,India**

Abstract

This paper is step to locate the researchers, their track in the way for recognizing the characters from various regional scripts in India. This survey paper would provide a path for developing recognition tools for Indian scripts where there is still a scope of recognition accuracy. It provides the reasons for researchers to work for recognition of Indian scripts. In this paper, the various scripts along with their special properties, there feature extraction and recognition techniques are described. Simultaneously a detailed comparison is made on the fronts of techniques used and recognition of Gujarati script.

Keywords: *Handwritten character recognition; feature extraction techniques; classification; Indian Scripts, Gujarati Script.*

1. Introduction

In the age of technological development each paper is somewhere in process to look forward to be in machine-editable format. Also automation of many official works has promoted the researchers to bridge up the gap between the common man and the technology.

Handwritten character recognition, usually abbreviated as HCR, is the recognition of handwritten text using computers. There are two types in which the characters are recognized: Offline and Online. As pen-paper was the prime way of communication the motivational factor branches us to offline character recognition. Various applications of offline handwritten character recognition are reading aid for the blind, preserving handwritten old/historical documents in electronic format, automatic reading for sorting of postal mail, bank cheques, atomization of various administrative offices, etc.

These days, vast research has been carried out for making commercially available efficient and inexpensive OCR packages to recognize printed texts. Printed characters can have variety of fonts and point sizes. A large amount of literature is available for the recognition of English, Japanese, Chinese, Arabian characters; whereas comparatively a meager amount of work has been reported for the recognition of Indian scripts [1, 5, 6, 8, 11, 14, 22, 28, 46]. There is variety in the handwritten Indian scripts on the fronts of basic consonants and vowels, there script-wise representation, there conjunctional appearance. The free-hand written characters itself is a challenge for recognition.

This survey paper would provide a path for developing recognition tools for Indian scripts where there is still a scope of recognition accuracy. In this paper, the various scripts along with their special properties, there feature extraction and recognition techniques are described. are described. Simultaneously a detailed comparison of Gujarati script is made on the fronts of techniques and recognition. This paper is a step for researchers to locate the track in the way for recognizing the characters from Gujarati script.

2. Steps for HCR

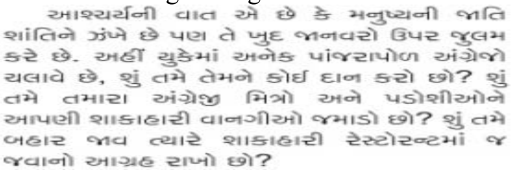
Whenever a document is thought for recognition, there are enumerable factors involved herewith. Firstly the document is scanned so that the text on paper becomes the image on computer. Then this image is preprocessed and then converted into either machine-editable format of just recognized as the set of characters or might be converted into some other script on PC as a language translation tool. To handle the image, preprocessing involves a lot of steps [3] so that the ratio for recognition enhances so also the motive of error reduction increases. These general preprocessing steps are summarized as under

- Binarization of scanned image
- Removal of Noise from scanned image
- Thinning of binarized image
- Skew detection and correction of scanned image,
- Segmentation of image
- Feature Extraction Techniques
- Recognition on the basis of Classifiers

2.1 Binarization

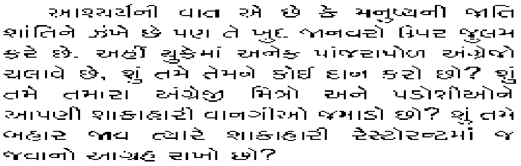
Binarization plays an important role in document processing. Due to binarization the segmentation of character and its recognition is affected. Basically separation of background and foreground of a scanned image is called binarization. The most popular technique for binarization is thresholding [9,48] in which an optimum threshold is selected and accordingly all the pixel intensities are converted to 1 i.e. background and 0 i.e. foreground.

The sample image Figure 1 is the simple scanned image from Gujarati News paper while after binarization it appears to be like image in Figure 2



આશ્ચર્યની વાત એ છે કે મનુષ્યની જાતિ
શાંતિને ઝંખે છે પણ તે ખુદ જાનવરો ઉપર જુલમ
કરે છે. અહીં યુકેમાં અનેક પાંજરાપોળ અંગ્રેજો
ચલાવે છે, શું તમે તેમને કોઈ દાન કરો છો? શું
તમે તમારા અંગ્રેજી મિત્રો અને પડોશીઓને
આપણી શાકાહારી વાનગીઓ જમાડો છો? શું તમે
બહાર જાવ ત્યારે શાકાહારી રેસ્ટોરન્ટમાં જ
જવાનો આગ્રહ રાખો છો?

Figure 1 Scanned image of Gujarati Samachar



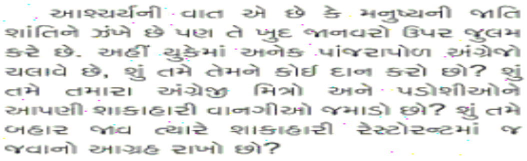
આશ્ચર્યની વાત એ છે કે મનુષ્યની જાતિ
શાંતિને ઝંખે છે પણ તે ખુદ જાનવરો ઉપર જુલમ
કરે છે. અહીં યુકેમાં અનેક પાંજરાપોળ અંગ્રેજો
ચલાવે છે, શું તમે તેમને કોઈ દાન કરો છો? શું
તમે તમારા અંગ્રેજી મિત્રો અને પડોશીઓને
આપણી શાકાહારી વાનગીઓ જમાડો છો? શું તમે
બહાર જાવ ત્યારે શાકાહારી રેસ્ટોરન્ટમાં જ
જવાનો આગ્રહ રાખો છો?

Figure 2 Binarized image of Gujarati Samachar (threshold level 0.4)

2.2 Removal of Noise

Digital images are prone to a variety of types of noise. Noise is the result of errors in the image acquisition process that result in pixel values that do not reflect the true intensities of the real scene. There are several ways that noise can be introduced into an image, depending on how the image is created. If the image is scanned from a photograph made on film, the film grain is a source of noise. Noise can also be the result of damage to the film, or be introduced by the scanner itself. If the image is acquired directly in a digital format, the mechanism for gathering the data (such as a CCD detector) can introduce noise. Electronic transmission of image data can introduce noise.

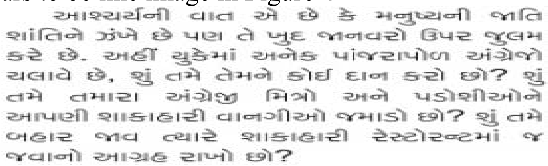
Whenever the image is created after scanning the document using any flat bed scanner there are chances of intrusion of some signals that are not the part of the image. To remove such excess signals is the process of noise removal which involves many filtering techniques [26].



આશ્ચર્યની વાત એ છે કે મનુષ્યની જાતિ
શાંતિને ઝંખે છે પણ તે ખુદ જાનવરો ઉપર જુલમ
કરે છે. અહીં યુકેમાં અનેક પાંજરાપોળ અંગ્રેજો
ચલાવે છે, શું તમે તેમને કોઈ દાન કરો છો? શું
તમે તમારા અંગ્રેજી મિત્રો અને પડોશીઓને
આપણી શાકાહારી વાનગીઓ જમાડો છો? શું તમે
બહાર જાવ ત્યારે શાકાહારી રેસ્ટોરન્ટમાં જ
જવાનો આગ્રહ રાખો છો?

Figure 3 Gujarat Samachar News paper with noise

The sample image Figure 3 is the image from Gujarati News paper with noise whereas after noise removal it appears to be like image in Figure 4



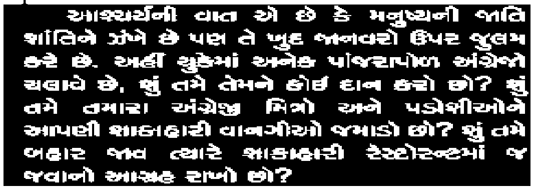
આશ્ચર્યની વાત એ છે કે મનુષ્યની જાતિ
શાંતિને ઝંખે છે પણ તે ખુદ જાનવરો ઉપર જુલમ
કરે છે. અહીં યુકેમાં અનેક પાંજરાપોળ અંગ્રેજો
ચલાવે છે, શું તમે તેમને કોઈ દાન કરો છો? શું
તમે તમારા અંગ્રેજી મિત્રો અને પડોશીઓને
આપણી શાકાહારી વાનગીઓ જમાડો છો? શું તમે
બહાર જાવ ત્યારે શાકાહારી રેસ્ટોરન્ટમાં જ
જવાનો આગ્રહ રાખો છો?

Figure 4 Gujarat Samachar News paper with noise removed

2.3 Thinning

Image thinning reduces a large amount of memory usage for structural information storage. Binary digital image can be represented by a matrix, where each element in matrix is either zero (white) or one (black) and the points are called pixels. Thinning is a process that deletes the unwanted pixels and transforms the image pattern one pixel thick .i.e. the thinning operation is typically applied repeatedly, leaving only pixel-wide linear representations of the image objects. The thinning operation halts when no more pixels can be removed from the image. This occurs when the thinning produces no change in the input image. At this point, the thinned image is identical to the input image [17].

The sample image Figure 5 is the binarized form of the image that was scanned while Figure 6 displays the results of the thinning operation, reducing the original objects to single pixel wide lines.



આશ્ચર્યની વાત એ છે કે મનુષ્યની જાતિ
શાંતિને ઝંખે છે પણ તે ખુદ જાનવરો ઉપર જુલમ
કરે છે. અહીં યુકેમાં અનેક પાંજરાપોળ અંગ્રેજો
ચલાવે છે, શું તમે તેમને કોઈ દાન કરો છો? શું
તમે તમારા અંગ્રેજી મિત્રો અને પડોશીઓને
આપણી શાકાહારી વાનગીઓ જમાડો છો? શું તમે
બહાર જાવ ત્યારે શાકાહારી રેસ્ટોરન્ટમાં જ
જવાનો આગ્રહ રાખો છો?

Figure 5 Binarized form of Gujarat Samachar News paper

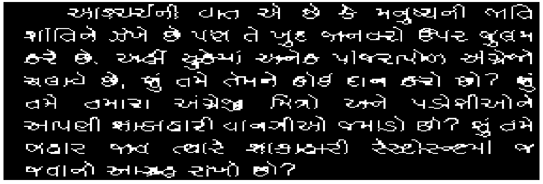

આશ્ચર્યની વાત એ છે કે મનુષ્યની જાતિ
શાંતિને ઝંખે છે પણ તે ખુદ જાનવરો ઉપર જુલમ
કરે છે. અહીં યુકેમાં અનેક પાંજરાપોળ અંગ્રેજો
ચલાવે છે, શું તમે તેમને કોઈ દાન કરો છો? શું
તમે તમારા અંગ્રેજ મિત્રો અને પડોશીઓને
આપણી શાકાહારી વાનગીઓ જમાડો છો? શું તમે
બહાર જાવ ત્યારે શાકાહારી રેસ્ટોરન્ટમાં જ
જવાનો આગ્રહ રાખો છો?

Figure 6 Thinned form of Gujarat Samachar News paper

2.4 Skew detection and correction

For any image created by scanning the document comprises of human intervention such as while feeding the paper for scanning the document it may be placed with some tilt in either of the direction. Or it may happen that while saving the scanned image slight rotation may occur due to human error or it may be the fact that the document itself may have handwritten characters with some angle made by human while writing the document.

Figure 7 shows the skew in the document and Figure 8 shows the resulted image after skew correction

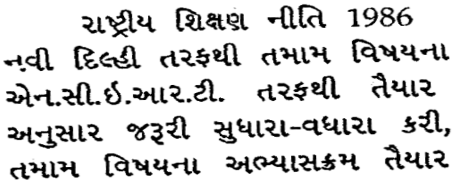

રાષ્ટ્રીય શિક્ષણ નીતિ 1986
નવી દિલ્હી તરફથી તમામ વિષયના
એન.સી.ઈ.આર.ટી. તરફથી તૈયાર
અનુસાર જરૂરી સુધારા-વધારા કરી,
તમામ વિષયના અભ્યાસક્રમ તૈયાર

Figure 7 Gujarati document having skew

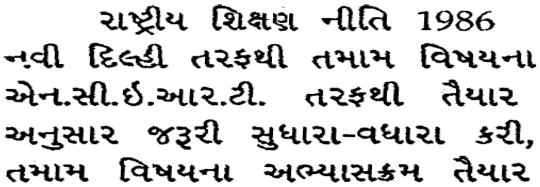

રાષ્ટ્રીય શિક્ષણ નીતિ 1986
નવી દિલ્હી તરફથી તમામ વિષયના
એન.સી.ઈ.આર.ટી. તરફથી તૈયાર
અનુસાર જરૂરી સુધારા-વધારા કરી,
તમામ વિષયના અભ્યાસક્રમ તૈયાર

Figure 8 Skew corrected image

The images in figure 7 and 8 are from the published work of Shah [31]

2.5 Segmentation of Image

The image cannot be handled completely at a glance. It has to be subdivided into many parts so that each part of the image is readable. To accomplish this task the image is subdivided considering three aspects, i.e. line wise segmentation, word wise segmentation and finally character wise segmentation.

While considering the line segmentation, the image is divided into the lines which make the understanding of image restricted to the lines in it and for doing so the algorithm works only for portioning the document image into small blocks called lines.

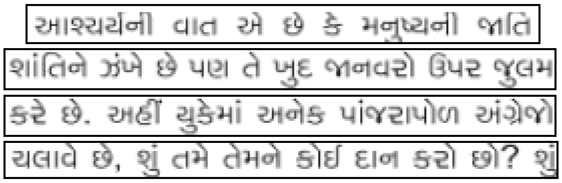

આશ્ચર્યની વાત એ છે કે મનુષ્યની જાતિ
શાંતિને ઝંખે છે પણ તે ખુદ જાનવરો ઉપર જુલમ
કરે છે. અહીં યુકેમાં અનેક પાંજરાપોળ અંગ્રેજો
ચલાવે છે, શું તમે તેમને કોઈ દાન કરો છો? શું

Figure 9 Line wise Segmentation of image from Gujarati Samachar

For considering the word wise segmentation, the image which is divided into the lines is further divided into words which now make the understanding of image restricted to the words in lines and for doing so the algorithm works only for portioning the document image into more small blocks called words

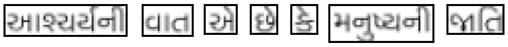

આશ્ચર્યની વાત એ છે કે મનુષ્યની જાતિ

Figure 10 Word wise Segmentation of image from Gujarati Samachar

For considering the character wise segmentation, the image which is divided into the lines is further divided into words is then further divided into characters which now make the understanding of image restricted to the characters in words from lines in documents and for doing so the algorithm works only for portioning the document image into more small blocks called characters

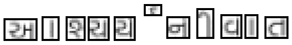

આશ્ચર્યની વાત

Figure 11 Character wise Segmentation of image from Gujarati Samachar

For doing all these types of segmentations many algorithms have been proposed [8]

2.6 Feature Extraction Techniques

Each and every character or numeral has some special and distinct parameters to represent and define them. But it also may happen that some of the parameters may collide while selecting the characters. So one needs to have such set of parameters in which each character is discriminated to its maximum extent. To find a set for parameters that defines the character is called feature extraction while subset of parameters which can define a character to its maximum extent is called as feature selection

The process of feature selection can be carried out at three fronts: Statistical Features, Syntactical/Structural Features and Hybrid Features. For statistical features, features are derived from statistical moments, geometrical moments, etc. For Syntactical/Structural features, features are derived as strokes, holes, end points, loops, cross-over or such structures in characters. The set of Hybrid features has the combination of Statistical and Structural features at necessary level of representation

2.7 Recognition on the basis of Classifiers

Once the features are selected, the step that then comes forward is for recognition. This process recognizes individual character and then results into the machine editable format. To perform this process many classifiers are available out of which some are very popular like template matching method or distance classifier like Euclidean distance measure.

Now-a-days nearest neighbor classifiers, fuzzy classifiers and Support Vector Machine classifiers are also claiming to give better results. Some researchers are using neural network as classifier.

3. Properties and Recognition Techniques of various Indian Scripts

Diversity in India, the scripts of India also show a wide range of variety in the characters and numerals of various scripts. One can trace the complexity among the Indian scripts. Many of the Indian regional scripts do not have shirorekha like Gujarati, Oriya, etc. Here for the study we have taken following prime scripts that are worked on:

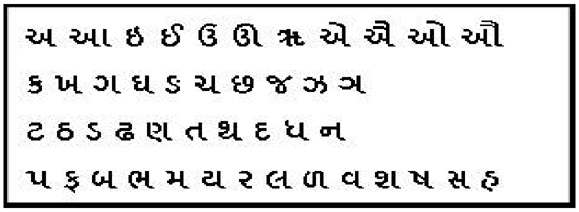
- Gujarati
- Devanagari
- Kannada
- Gurumukhi
- Oriya
- Tamil
- Telugu
- Bengali

3.1 Gujarati Script

The Gujarati script was adapted from the Devanagari script to write the Gujarati language. The earliest known document in the Gujarati script is a manuscript dating from 1592, and the script first appeared in print in a 1797 advertisement. Until the 19th century it was used mainly for writing letters and keeping accounts, while the Devanagari script was used for literature and academic writings.

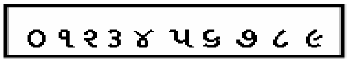
Gujarati, an Indo-Aryan language spoken by about 46 million people in the Indian states of Gujarat, Maharashtra, Rajasthan, Karnataka and Madhya Pradesh, and also in Bangladesh, Fiji, Kenya, Malawi, Mauritius, Oman, Pakistan, Réunion, Singapore, South Africa, Tanzania, Uganda, United Kingdom, USA, Zambia and Zimbabwe. The major difference between Gujarati and Devanagari is the lack of the top horizontal bar in Gujarati. Otherwise the two scripts are fairly similar [35].

The basic consonants and vowels of the Gujarati script are shown in Figure 12.



અ આ ઇ ઈ ઉ ઊ ઋ એ ઐ ઓ ઔ
ક ખ ગ ઘ ઙ ચ છ જ ઝ ઞ
ટ ઠ ડ ઢ ણ ત થ દ ધ ન
પ ફ બ ભ મ ય ર લ ળ વ શ ષ સ હ

Figure 12 Vowels and Consonants of Gujarati Script



૦ ૧ ૨ ૩ ૪ ૫ ૬ ૭ ૮ ૯

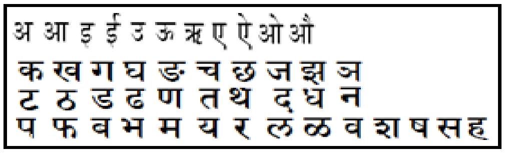
Figure 13 Numerals of Gujarati Script

Figure 13 shows the ten numerals from Gujarati script

3.2 Devanagari Script

More than 500 million people speak and write Devanagari script around the world. Many languages use Devanagari script [32] like Hindi and Marathi. Devanagari has 11 vowels and 33 simple consonants. Besides the consonants and the vowels, other constituent symbols in Devanagari are set of vowel modifiers called matra (placed to the left, right, above, or at the bottom of a character or conjunct), pure-consonant (also called half-letters) which when combined with other consonants yield conjuncts. A horizontal line called shirorekha (a header line) runs through the entire span of work [45].

The figure 14 shows the basic Devanagari alphabets while figure 15 shows the numerals of Devanagari script.



अ आ इ ई उ ऊ ऋ ए ऐ ओ औ
क ख ग घ ङ च छ ज झ ञ
ट ठ ड ढ ण त थ द ध न
प फ ब भ म य र ल ळ व श ष स ह

Figure 14 Basic vowels and consonants of Devanagari

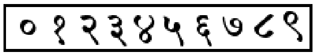


Figure 15Numerals in Devanagari

A lot of work has been reported for printed Devanagari text, whereas very little is reported for handwritten Devanagari script. For the first time Sethi [34] worked for hand printed characters and for typed Devanagari script Sinha and Mahabala put their efforts [27]. Sinha and Bansal [36] achieved 93% performance on individual characters. Recognition of Devanagari text in Sanskrit manuscript 'Saddharmapundarika' [37] is achieved with an accuracy of 98.09% using structural features and neural networks for classification.

Pal and Chaudhuri have attempted OCR for two scripts, Bangla and Devanagari in [38]. Database evaluation methods are given in [39] and database for Devanagari numerals has been collected from mail addresses and job application forms in [40]. Machine recognition of online handwritten Devanagari characters has been reported in [33] with 82-85% accuracy. In [41] online Devanagari script recognition is attempted with 86.5% accuracy on a database of 20 writers. A combination of on-line and offline features has been used in [42] Binary Wavelet transform is used for feature extraction of handwritten Devanagari characters. In [43], a survey of different structural techniques used for feature extraction in OCR of different scripts is given. Recently in [44], Quadratic classifier based method is proposed with 81% accuracy.

3.3 Kannada Script

The Kannada script is used in the southern Indian state of Karnataka to write the Kannada language. It is derived from the Old Kannada script and is closely related to the Telugu script. The components of Kannada script are shown in figure 16 and numerals in figure 17 respectively.

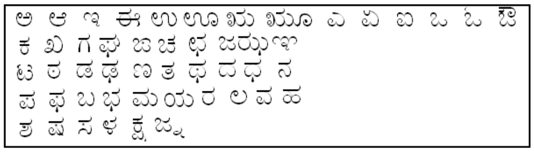


Figure 16 Components of Kannada Script

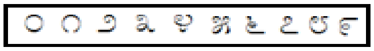


Figure 17Numerals of Kannada Script

In [8] the author split each Kannada segment image into number of zones in the radial and the angular directions extracting the features capturing the shape of the characters and used Support Vector Machine as the classifier and achieved 87 to 95% accuracy

3.4 Gurmukhi Script

Gurmukhi script alphabets consist of 41 consonants and 12 vowels as shown in figure 18. It also contains 10 numerals as shown in figure 19. Besides these, some of the characters in form of half characters are present in the feet of characters. Writing style is from left to right. The concept of upper/lowercase characters is absent in Gurmukhi. A line of Gurmukhi script may be partitioned into three horizontal zones, the middle zone being the busiest one. The upper and lower zones may contain parts of vowel modifiers and diacritical markers.

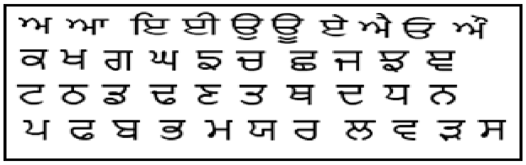


Figure 18Basic alphabets of Gurumukhi Script

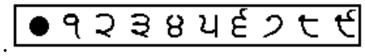


Figure 19 Numerals in Gurumukhi Script

Lehal and Singh presented an OCR system for printed Gurumukhi script [46]. The skew angle is determined by calculating horizontal and vertical projections at different angles at fixed interval in the range [0° to 90°].A recognition rate of 96.6% at a processing speed of 175 characters/second was reported. Lehal and Singh [16] also developed a post processor for Gurmukhi.

3.5 Oriya Script

The Oriya script developed from an early form of the Bengali script, which belongs to the Northern group of South Asian scripts. Oriya is used to write the Oriya language, which is spoken in the modern Indian state of Orissa, located on the east coast of India. While the cursive shapes of the Oriya letters appear to suggest influences from Southern scripts, it is thought that the cursive shape evolved from the need to write on palm leaves with a pointed stylus, which has a tendency to tear if you use too many straight lines. One can notice from figure 20, the round form of alphabets used in Oriya script. Figure 21 shows the numerals used in Oriya script.

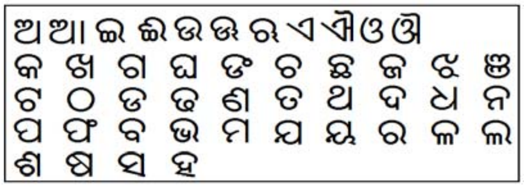


Figure 20 Oriya letters

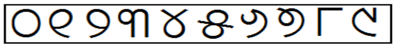


Figure 21Oriya numerals

Chaudhari et al.[5] used preprocessing techniques like skew correction, line segmentation, zone detection, word and character segmentation and then the combination of stroke and run-number based and water reservoir based features were used as classifiers. They achieved 96.3% of accuracy.

3.6 Tamil Script

Tamil is a Dravidian language and one of the oldest languages in the world. It is the official language of the Indian state of Tamil Nadu; it also has official status in Sri Lanka, Malaysia and Singapore. The Tamil script has 10 numerals, 12 vowels, 18 consonants (as shown in figure 22 and 23) and five grantha letters. The script, however, is syllabic and not alphabetic. The complete script, therefore, consists of 31 letters in their independent form, and an additional 216 combining letters representing every possible combination of a vowel and a consonant.

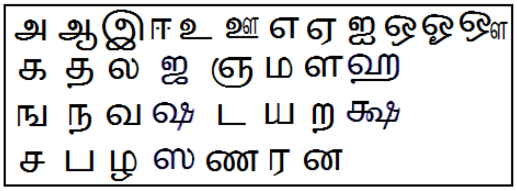


Figure 22 Tamil Letters

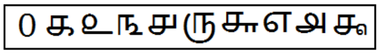


Figure 23Tamil Numerals

Siromony et al. [28] described a method for recognition of machine printed letters of the Tamil alphabet using an encoded character string dictionary.

3.7 Telugu Script

Telugu is one of the ancient (5000 years old) languages of India with rich cultural heritage. This language is a mother tongue of 100 million population in southern part of India. It is an abugida from the Brahmic family of scripts. It has a complex orthography with a large number of distinct character shapes composed of simple and compound characters. The basic alphabet of Telugu consists of 16 vowels (called achchus) and 36 consonants (called hallus) totaling to 52 symbols as shown in figure 24. Each vowel in Telugu is associated with a vowel signs.

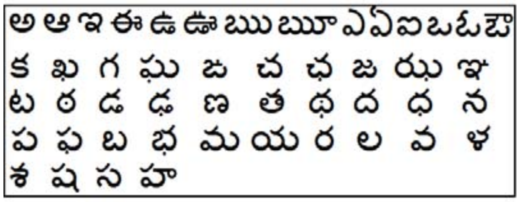


Figure 24Telugu vowels and consonants

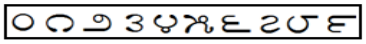


Figure 25Telugu numerals

The first reported work on OCR of Telugu Character is by Rajasekaran and Deekshatulu [25]. Sukhaswami et al. [29] proposed a neural network based system. Hopfield model of neural network working as an associative memory is chosen for recognition purposes initially.

Pujari et al. [24] used wavelet multi-resolution analysis for capturing the distinctive characteristics of Telugu script and associative memory model for recognizing the characters. The author had very conservative recognition rate across fonts and sizes and is reported as varying from 93% to 95%. An OCR for Telugu is reported by Negi, et al.. [18]. Raw OCR accuracy with no post processing is reported as 92%. Performance across fonts varied from 97.3% for Hemalatha font to 70.1% for the newspaper font. Non-linear normalization to improve performance was used by Negi et al.., [19] by selectively scaling regions of low curvature in the glyphs. Negi and Nikhil [20] attempted Layout analysis to locate, and extract Telugu text regions from document images. The gradient magnitude of the image was computed to obtain contrasting regions in the image. Hough Transform for circles was applied on the gradient magnitude of the image to obtain the circular gradient which is a prominent feature of Telugu text. Each detected circle is filled to obtain the regions of interest. Recursive XY cuts and projection profiles are used to segment the document image into paragraphs, lines, and words.

Lakshmi and Patvardhan [15] presented recognition of basic Telugu symbols. Feature vector is computed out of a set of seven invariant moments from the second and third order moments. Recognition is done using k-nearest neighbor algorithm on these feature vectors. Jawahar et al. [14] proposed a Bilingual OCR for Hindi-Telugu documents. It is based on Principal Component Analysis followed by support vector classification. An overall accuracy of approximately 96.7% is reported.

Anuradha Srinivas, et al. [4] developed a Telugu optical character recognition system for a single font. A 2-stage classifier with first stage identifies the group number of the test character, and a minimum-distance classifier at the second stage identifies the character. Recognition accuracy of 93.2% is reported.
In the reported work by Pratap Reddy et al. [33] structural features of the syllable and the component model are combined to extract middle zone components. The shape of the middle zone components is closely related to a circle whereas other components are found with different topological features. Recognition rate of 99 percent is observed with the proposed method.

3.8 Bengali Script

Bengali is a Nagari-derived script that appeared in eastern South Asia around the 11th century CE. It is still currently used in Bangladesh, as well as the state of West Bengal in India (hence the script's name) on the eastern part of India. The old Bengali script (11th century CE) is also the parent to many other scripts of eastern India, such Oriya, Manipuri, and Maithili. The Bengali script is used to writer languages in eastern India such as Bengali, Assamese, and Manipuri.

Basic Bengali character set comprises 11 vowels, 39 consonant, 10 numerals as shown in figure 26 and 27. There are also compound characters being combination of consonant with consonant as well as consonant with vowel. A vowel following a consonant sometimes takes a modified shape and is called a vowel modifier. Many characters of Bengali script have a horizontal line at the upper part called ‘matra’ or headline.

.

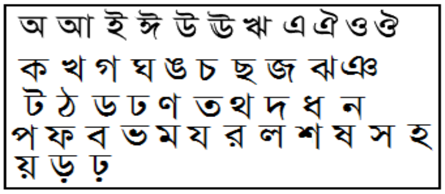


Figure 26 Bengali Character set

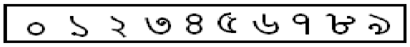


Figure 27Bengali Numerals

Recognition of isolated and continuous printed multi font Bengali characters is reported in the work by Mahmud et al. [13].Using chain code representation, classification is done by a feed forward neural network. Testing on three types of fonts with accuracy of approximately 98% for isolated characters and 96% for continuous characters is reported. A complete OCR for printed Bangla is reported in the work by Chaudhari and Pal [11], in which a combination of template and feature-matching approach is used. For single font, clear documents 99.10% character level recognition accuracy is reported
In the reported work of Shamik et al. [47] fuzzy features are extracted from Hough transform of a character pattern pixels MLPs used for classification have the fuzzy features as inputs. During recognition, the class of each pattern is first determined, followed by recognition of the actual character within that class. Recognition accuracy of the system is more than 98%.

4. Recognition Techniques used for Gujarati Script

Methods and recognition rates depend on the level of constraints on handwriting. The constraints are mainly characterized by the types of handwriting, the number of scripter, the size of the vocabulary and the spatial layout. Obviously, recognition becomes more difficult when the constraints decrease. For Gujarati scripts the reported contribution is described as below.

Antani and Agnihotri [1] in 1999 have given the primitive effort to Gujarati printed text. The author has created the data sets from scanned images, at 100 dpi, of printed Gujarati text as well as from various sites of internet from 15 font families. For training 5 fonts created 10 samples each. The images were scaled up and then scaled down to a fixed size so that all the samples should be of same size i.e. 30x20. It does not have skew correction or noise removal etc. for feature extraction the author computed both invariant moments and raw moments. Also image pixel values are used as features creating 30x20= 600 dimensional binary feature space. For classification the author has used two classifiers, K-NN classifier and minimum hamming distance classifier. The best recognition rate was for 1-NN for 600 dimensional binary features space i.e. 67% 1-NN in regular moment space gave 48% while minimum distance classifier had the recognition rate of 39%. The Euclidean minimum distance classifier recognized only 41.33%.

Dholakia [6] attempted to use wavelet features, GRNN classifier and KNN classifier on the printed Gujarati text of font sizes 11 to 15 with styles regular, bold and italic by finding the confusing sets of the characters. They collected 4173 samples of middle zone glyphs of initial size 32x32 and 16x16 wavelet coefficients have been extracted creating the feature vector. Two sets of the randomly selected glyphs (2802 symbols) were used for training and 1371 symbols were used for testing. Two classifiers GRNN and KNN with Euclidean distance as similarity measure are used producing 97.59 and 96.71 as their respective recognition rates.

In 2005, Jignesh Dholakia et. al [7] have presented an algorithm to identify various zones used for Gujarati printed text. In the algorithm they have proposed the use of horizontal and vertical profiles. They have identified these zones by slope of lines created by upper left corner of rectangle created by the boundaries of connected components from line level and not word level the 3 different document images, 20 lines were extracted where 19 were detected with correct zone boundary. The line where it failed was very much skewed.

Desai [2] collected 300 samples of 300dpi with initial size of each numeral as 90x90. the author then adjusted the contrast by CLAHE i.e. contrast limited adaptive histogram equalization algorithm considering 8x8 tiles and 0.01 as contrast enhancement constant. The boundaries are then smoothed out by median filter of 3x3 neighborhoods. Image is then reconstructed to the size of 16x16 pixels using nearest neighbor interpolation.

For feature extraction four profile vectors are used as an abstracted feature of identification of digit. Five more patterns for each digit are created in both clockwise and anticlockwise directions with the difference of 2degrees each up to 10°. A feed forward back propagation neural network is used for Gujarati numeral classification with 278 sets of various digits. Out of these 278 sets, 11 sets were created by a standard font. From the 265 sets the author recorded the success rate for standard fonts as 71.82%, for handwritten training sets as 91.0% while for testing sets as a score of 81.5% was recorded.

Another effort contributed for Gujarati script was by Shah & Sharma[31]. They used template matching and Fringe distance classifier as distance measure. Initially the sample images were filtered using low pass filter. Then the binarization was done by considering the optimal threshold method. Skew detection and correction is done within 0.05°. They segmented printed characters in terms of lines, words and connected components. By this effort, for connected component recognition rate was 78.34%. for upper modifier recognition rate was 50% where as for lower modifier it was 77.55% and for punctuation marks it was 29.6%. Cumulative for overall it was 72.3%.

The characteristics which constrain hand writing may be combined in order to define handwriting categories for which the results of automatic processing are satisfactory.

5. Conclusions

A study is made on different feature extraction and recognition techniques used for Indian scripts. In this paper, the different properties of the scripts, there components and the methods used to segment and identify the characters or scripts are being elaborated. Moreover the detailed study of Gujarati script is done which can be used as a starting step for the researchers entering into this area.

References

[1] S. Antani, L. Agnihotri. “Gujarati Character Recognition”. Fifth International Conference on Document Analysis and Recognition (ICDAR'99), 1999 p. 418.

[2] A. Desai, “Gujarati handwritten numeral optical character reorganization through neural network”. Pattern Recognition Vol 43 2010 pp. 2582-2589

[3] Anbumani, Subramanian. “Optical Character Recognition of Printed Tamil Characters”. Department of Electrical and computer Engineering, Virginia Tech, Blacksburg 2000

[4]Anuradha Srinivas, A. Agarwal, C.R.Rao “Telugu Character Recognition”. Proc. Of International conference on systemics, cybernetics, and informatics, Hyderabad, 2007 pgs.654-659.

[5] B. B. Chaudhari, U. Pal and M. Mitra “Automatic recognition of printed Oriya script” Sadhana Vol 27(1) 200) pp23-34

[6] J. Dholakia, A. Yajnik, A. Negi “Wavelet Feature Based Confusion Character Sets for Gujarati Script”, ICCIMA, 2007 p 366-371

[7] J. Dholakia, A. Negi, S. Rama Mohan “Zone Identification in the Printed Gujarati Text”, Proc. of 8th ICDAR, 2005 p272-276

[8] T V Ashwin and P S Sastry “A font and size independent OCR system for printed Kannada documents using support vector machine”. Saadhanaa, Vol. 27, Part 1, 2002, pp. 35–58.

[9] B Anuradha, B Koteswarrao “An efficient Binarization technique for old documents.” Proc. Of International conference on Systemics, Cybernetics, and Inforrmatics (ICSCI2006), Hyderabad, 2006 pp771-775.

[10] B.B. Chaudhuri and U. Pal “Skew Angle Detection of Digitized Indian Script Documents.” IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 19, NO. 2, 1997 pp.

[11] B.B. Chaudhuri and U. Pal, “A complete printed Bangla OCR system”. Pattern Recognition, Vol. 31, 1998, pp 531-549.

[12] R. C. Gonzalez, and R. E. Woods Digital Image Processing. (New Jersey: Prentice-Hall) 2002

[13]Jalal Uddin Mahmud, Mohammed Feroz Raihan and Chowdhury Mofizur Rahman “A Complete OCR System for Continuous Bengali Characters”. Conference on Convergent Technologies for Asia-Pacific Region (TENCON) Volume 4, Issue, 15-17 2003 Page(s): 1372 – 1376

[14] Jawahar C. V., M. N. S. S. K. Pavan Kumar, S. S. Ravi Kiran "A Bilingual OCR for Hindi-Telugu Documents and its Applications." International Conference on Document Analysis and Recognition 2003.

[15] Lakshmi C V, C Patvardhan, Optical Character Recognition of Basic Symbols in Printed Telugu Text. IE(I)Journal-CP Vol 84, 2003 pgs.66-71.

[16] Lehal G S and Chandan Singh, A post-processor for Gurmukhi OCR Saadhana Vol. 27, Part 1, 2002, pp.99–111.

[17] Lam, L., Seong-Whan Lee, and Ching Y. Suen "Thinning Methodologies-A Comprehensive Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 14, No. 9, 1992 page 879

[18] A. Negi, B. Chakravarthy and Krishna B "An OCR system for Telugu." Proc. Of 6th Int. Conf. on Document Analysis and Recognition IEEE Comp. Soc. Press, USA, 2001. Pgs. 1110-1114.

[19] A. Negi, B. Chakravarthy and V.V.Suresh Kumar. "Non-linear Normalization to Improve Telugu OCR" Proc. of Indo-European Conf. on Multilingual Communication Technologies, Tata McGraw Hill Book Co., New Delhi, 2002 pgs 45-57.

[20] A. Negi, N. Kasinadhuni "Localization and Extraction of Text in Telugu Document Images" Conference on Convergent Technologies for Asia-Pacific Region (TENCON) 2003 pgs. 749-752

[21] U. Pal, B.B. Chaudhuri "Indian script character recognition: a survey." Pattern Recognition 37 2004 pgs.1887 –1899

[22] U. Pal, B.B. Chaudhuri "Printed Devanagari Script OCR System". Vivek, Vol.10, 1997 pgs.12-24

[23]P. Iyer, A. Singh, S. Sanyal "Optical Character Recognition System for Noisy Images in Devnagari Script." UDL Workshop on Optical Character Recognition with Workflow and Document Summarization 2005

[24] A. K Pujari , C Dhanunjaya Naidu & B C Jinaga "An Adaptive Character Recognizer for Telugu Scripts using Multiresolution Analysis and Associative Memory". ICVGIP, Ahmedabad. 2002

[25] S. N. S. Rajasekaran, B. L Deekshatulu. Recognition of printed Telugu characters. Comput. Graphics Image Processing, 1977 pgs.335–360.

[26] K. Rangachar, Lawrence O'Gorman and V. Govindaraju "Document image analysis: A primer." Saadhanaa Vol. 27, Part 1, 2002 pp. 3–22.

[27] R. M. K. Sinha and H. N. Mahabala "Machine recognition of Devnagari script." IEEE Trans. Systems Man Cybern 1979. Pgs.435–441.

[28] G. Siromony, R. Chandrasekaran, M. Chandrasekaran "Computer recognition of printed Tamil characters." Pattern Recognition Vol.10 1978 pgs.243–247.

[29] R Sukhaswami, P. Seetharamulu, A. K. Pujari "Recognition of Telugu characters using Neural networks", Int. J. Neural Syst. Vol.6 1995 pgs. 317–357.

[30] V. Bansal "Integrating knowledge sources in Devnagari text recognition". Ph.D. Thesis, IIT Kanpur, 1999

[31] S K Shah and A Sharma "Design and Implementation of Optical Character Recognition System to Recognize Gujarati Script using Template Matching" IE(I) Journal–ET Vol.86 2006 pgs. 44-49

[32] S. Kompalli, S. Kayak, S. Setlur and V. Govindaraju, "Challenges in OCR of Devnagari Documents" Proceedings of Eighth International Conference on Document Analysis and Recognition (ICDAR'05), Seoul, South Korea 2005. pp. 327-331.

[33] Pratap Reddy L., Satyaprasad L., A.S.C.S. Sastry, "Middle Zone Component Extraction and Recognition of Telugu Document Image," icdar, vol. 2, 2007 pp.584-588,

[34] I.K. Sethi, "Machine Recognition of Online Handwritten Devnagari Characters" Pattern Recognition, Vol. 9, 1977 pp. 69 - 75.

[35] http//en.wikipedia.org Accessed 15 June 2007

[36] R. M. K. Sinha and Veena Bansal, "A complete OCR for Printed Hindi Text in Devnagari Script". Proceedings of Fifth International Conference on Document Analysis and Recognition, Seattle, USA, September, 2001, pp. 800 - 804.

[37] K. Keeni, T-Nishino, H. Shimodaria and Y. Tan, "Recognition of Devnagari characters using Neural Networks." IEICE Transactions on Information and Systems, Vol. E79-D, No.5, pp. 523 -528, May, 1996.

[38] B.B. Chaudhuri and U. Pal," An OCR system to read two Indian Language Scripts: Bangla and Devnagari(Hindi)." Proceedings of Fourth International Conference on Document Analysis and Recognition, Ulm, Germany, August, 1997.pp. 1011 - 1015,

[39] S. Kompalli, S. Setlur, V. Govindaraju and R. Vemulapati, "Creation of data resources and design of an evaluation test bed for Devnagari script recognition". Proceedings of the thirteenth International Workshop Research Issues on Data Engineering: Multi- lingual Information Management Hyderabad, India, March, 2003, pp. 55 - 61,.

[40] U. Bhattacharya and B. B. Chaudhuri, "Databases for Research on Recognition of Handwritten Characters of Indian Scripts". Proceedings of Eighth International Conference on Document Analysis and Recognition,, Seoul, South Korea, September, 2005. pp. 789 – 793.

[41] S. D. Connel, R.M.K. Sinha, A. K. Jain, "Recognition of Un-constrained On-Line Devnagari Characters". Proceedings of ICPR, , Barcelona, Spain, September, 2000 pp. 2368 - 2371.

[42] P. Mukherji, V. B. Gapchup and Priti P. Rege, "Feature Extraction of Devnagari Characters using sub bands of Binary Wavelet Transform" Proceedings of RETIS-06, Kolkata, India,July,2006 pp. 176 -179.

[43] P. Mukherji and P. P. Rege. "A Survey of Techniques for Optical Character Recognition of Handwritten Documents with reference to Devnagari Script." Proceedings of First International Conference on Signal and Image Processing Hubli, India, December,2006, pp. 178 - 184,.

[44] N. Sharma, U. Pal, F. Kimura and S. Pal, "Recognition of Off-Line Handwritten Devnagari Characters Using Quadratic Classifier." Proc. of Indian Conference on Computer Vision Graphics and Image Processing (ICVGIP), India, pp. 805 - 816, Madurai, India, December,2006.

[45] V. Bansal and R. M. K. Sinha, "Integrating knowledge Sources in Devnagri Text Recognition," IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 30, July 2000, pp. 500 – 505.

[46] G S Lehal, C Singh "A Gurmukhi script recognition system". Proc. 15th Int. Conf. on Pattern Recognition (Los Alamitos, CA: IEEE Comput. Soc.) vol. 2, 2000 pp 557–560

[47] Shamik Sural P.K.Das, "Recognition of an Indian Script using Multilayer Perceptrons and Fuzzy Features" 2001 IEEE 1120-1124

[48] Otsu, N A "Threshold Selection Method from Gray-Level Histograms", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 9, No. 1, 1979 pp. 62-66.

Mamta Maloo has received M.Sc. Degree in Computer Science from Sant Gadge Baba Amravati University in 2003. Since 2007, she has been a Ph.D. student of Dr. K. V. Kale. Her current research interests include pattern recognition, image analysis and document processing.

Dr. Karbhari Vishwanath Kale, M.Sc., MCA (Engg & Tech), Ph.D., FIETE., Professor and Head Dept. of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada, University, Aurangabad, INDIA Completed M.Sc Physics (Electronics) in 1987, B.Ed in 1989, MCA (Engg. And Tech) in 1995 and Ph.D on Superionics in 1997. Fifteen students awarded Ph. D. in Computer Science. Ten students have working for Ph.D. under the guidance. One student awarded and two are working for M.Phil. under the guidance