2 LITERATURE REVIEW FOR HANDWRITTEN TEXT RECOGNITION

The objective of HTR system is to recognize handwritten text from a digitized document. The general architecture of HTR as given in topic 1.3, consists of preprocessing, segmentation, features extraction and recognition phases. The HTR system differs from isolated HCR as it requires segmentation phase. It is essential and challenging task for recognition of document containing text [57,58].

The overall approach of text recognition i.e. holistic or analytical [13,26,54,55,62], decides output of segmentation which is line, word, or character. The output of segmentation then processed for recognition and generation of text. The literate review on segmentation techniques are presented followed by review on features & classifiers. Later in this chapter major work on handwritten Gujarati OCR is presented.

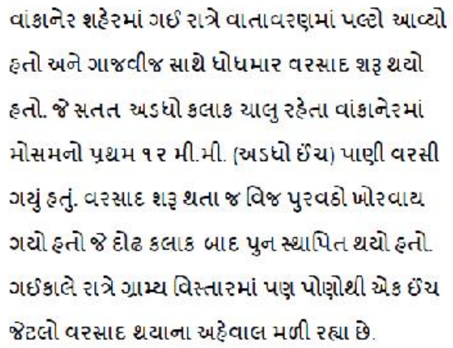
2.1 LITERATURE REVIEW ON SEGMENTATION

Numbers of techniques have been proposed for segmentation of text document image. The segmentation technique is selected based on various factors like overall text recognition approach, language characteristics, nature of writing style, complexity of character shapes etc.

The text document contains multiple lines hence line segmentation is key requirement of segmentation phases. Most popular way of segmenting line is using horizontal projection [89]. The projection $H_{[i]}$ along the row of binary image is given by [65]:

$$H_{[i]} = \sum_{i=1}^{n} f_{(i,j)} \qquad \text{Equation 1}$$

The sample printed binary text document shown in Figure 5 and its histogram is shown in Figure 6. In horizontal projection the mountain region represents the presence of text lines, while valley represents gap between text lines. By identifying valley points from the text document, line segmentation points are obtained.



વાંકાનેર શહેરમાં ગઈ રાત્રે વાતાવરણમાં પલ્ટો આવ્યો
હતો અને ગાજવીજ સાથે ધોધમાર વરસાદ શરૂ થયો
હતો. જે સતત અડધો કલાક ચાલુ રહેતા વાંકાનેરમાં
મોસમનો પ્રથમ ૧૨ મી.મી. (અડધો ઈંચ) પાણી વરસી
ગયું હતું. વરસાદ શરૂ થતા જ વિજ પુરવઠો ખોરવાય
ગયો હતો જે દોઢ કલાક બાદ પુન સ્થાપિત થયો હતો.
ગઈકાલે રાત્રે ગ્રામ્ય વિસ્તારમાં પણ પોણોથી એક ઈંચ
જેટલો વરસાદ થયાના અહેવાલ મળી રહ્યા છે.

Figure 5 Handwritten printed text document

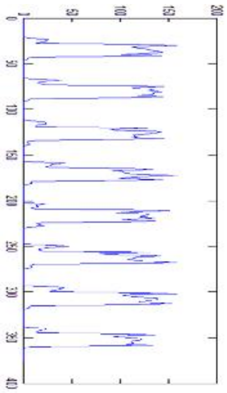


Figure 6 Horizontal projection

The horizontal projection is commonly used approach for line segmentation of printed document and used for properly handwritten document with little overlap. However, in handwritten documents the irregularity of writing causes the internal skew and due to that projection method is not suitable [90]. Variation of projection based method used by many researchers for segmentation of handwritten document [75,89,91].

In [61,92,93], grouping method was used, which is bottom up approach for segmentation based on aggregating unit which is pixel or higher level to make alignments such as connected component or block. The units are then joined to form alignments. The joining schemes are based on local and global criteria. The connected components are connected to close one based on neighbour-hood or geometrical criteria to form text lines. The shortcoming of grouping methods are 1) need for initial alignment, 2) defining criteria for reaching to next unit, and 3) solving conflict if unit belongs to more than one alignments.

In [37], combination approach was used for segmentation of touching, overlapping, skewed and short lines in handwritten English text. The non-overlapping lines segmented using horizontal projection profile. The lines which do not share pixels with other lines are segmented using connected components. To solve issue of touching and overlapping line run-length is used to cut and assign component to appropriate line. To handle skew lines the component assigned to line using mapping distance matrix and to identify short line right profile is used. The segmentation algorithm tested on randomly selected samples of IAM dataset and reported 91.92% of line segmentation accuracy.

In [74,94,95], segmentation method based on Hough transform which considers any image to compose of straight lines was used. It creates an angle, offset plane in which the local maxima are assumed to correlate with text lines. In [94], minima point of connected component was used as input to Hough transform to detect line. In [74], centroid of connected component was used as input to Hough transform to detect set of connected components lies on straight line.

The run length smoothing algorithm (RLSA) is applied on binary image to produce the effect of linking together neighbouring area producing connected text region. The Figure 7 shows binary image after applying RLSA on image which is shown in Figure 5. Using connected component, text region is extracted and it may contain line or word as single component which is based on run-length.

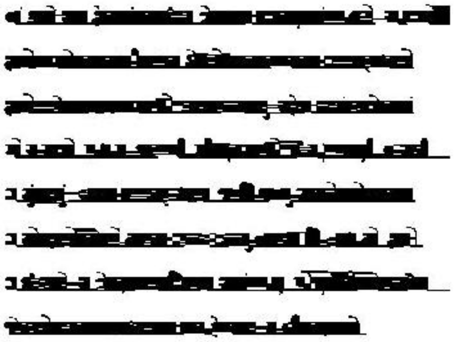


Figure 7 Binary image after RLSA

In [12], line segmentation method based on smearing and connected component was used for extracting text region from the printed document. In [96], fuzzy run length approach used to group line patterns for historical document containing handwritten text written in English script. Using the threshold value of run length, connected component selected as text lines.

In [97], a technique based on piecewise projection profile and connected component was used for English, French, German and Greek scripts handwritten text line segmentation and word segmentation. The document image is divided into vertical zones and extreme points of piecewise projection profiles are used to over-segment each zone in text and gap region. Then the optimal succession point of text and gap areas within vertical zones obtained by applying Viterbi algorithm. The connected components are assigned to text line after applying text line separator drawing technique. The word segmentation is based on a gap metric that exploits the objective function of a soft-margin linear SVM that separates successive connected components. The algorithms tested on the benchmarking datasets of ICDAR07 handwriting segmentation contest and achieved 98.33% for lines and 93.01% for words.

In [93], shredding technique was used to extract text lines from the document by shredding their surface with local minima tracers. The approach is based on the topological assumption that for each text line, a path exists from one side of the image to the other that traverses only one text line. They first blur the image and then use tracers to follow the white-most and black-most paths from left to right as well as from right to left in order to shred the image into text line areas. After tracing line areas (white run), line centres where assigned labelled. Shredding technique achieves 98.6% recognition rate on ICDAR07 dataset for English, French, Germen and Greek scripts, which is better compare to [97]. The said technique uses gray scale image while in [97] uses operation on binary image.

In [90] used the method based on perspective vision to segment text lines and words for English, Greek, French and German documents of ICDAR'09 dataset. For the detection of text lines, perceptive vision that is global vision of document (at low resolution) enables the detection of text line position as they were line segments. The vision in high resolution enables confirmation of presence of text lines. Then precise assignment of each pixel to each text line is carried out using re-segmentation of connected component. The segmentation algorithm tested on set of ICDAR’2009 obtained 99.25% of accuracy for lines and 94.20% of accuracy for words.

In [98] used a comprehensive algorithm for offline handwritten Arabic script segmentation based on vertical and horizontal histogram. The segmentation technique uses the vertical histogram to segment Arabic word into horizontal character primitives. Then process followed by segmenting horizontal character primitives into vertical character primitives using horizontal histogram. By tracing upper contour, the order of character is determined.

In [61], connected component approach was used for segmenting words in handwritten Arabic text. In this method firstly, connected components are extracted, and distances among different components are analysed. The statistical distribution of this distance is then obtained to determine an optimal threshold for words segmentation. An improved projection based method is also employed for baseline detection.

In [18], morphological method, zone projection and character separation by vertical projection was used for segmentation of Japanese handwritten text. The morphological closing and strip wise horizontal projection profile was used to extract text line. For character segmentation again, morphological closing and strip wise vertical projection profile was used.

All above segmentation strategies mainly used for document containing foreign languages like English, French, German, Greek, Arabic and Japanese scripts. Major challenges in these scripts are with respect to skew lines, and touching lines. Other challenges are with respect to cursive writing which does not separate characters. Also, to note that majority segmentation approach uses combination approach based on projection histogram, run-length and connected component and gives better results.

Next discussion will be on segmentation techniques which were used for Indic script. The major work founds in text recognition concentrating script Devnagari, Gurmukhi, Oriya, Bangla, Assamese, Marathi and Kannad.

In [22], a piece-wise projection approach was used for segmentation of handwritten text document written in unconstrained Bangla. The width of strip calculated based on statistical analysis. Using piece-wise projection piece wise separating lines were detected and joined. The vertical projection is used to segment word from the lines for extracting characters from the word water reservoir concept is used. After identification of isolated character touching character segmented using base area points of water reservoir.

In [23], same technique was used with modification in word and character segmentation for unconstrained Oriya handwritten text. For line segmentation, the document is divided into vertical stripes and piece-wise projection is obtained. By analysing the heights of the water reservoirs obtained from different components of the document, the width of a stripe is calculated. Stripe-wise horizontal histograms are then computed and the relationship of the peak–valley points of the histograms is used for line segmentation. The segmentation algorithm achieved more than 97% of accuracy in line segmentation. Based on vertical projection profiles and structural features of Oriya characters, text lines are segmented into words.

In [99], ANN based segmentation method was used for Assamese handwritten text. In Assamese the presence of head line helps in segmentation but without head line characters appearing similar and making recognition difficult. The initial line and character segmentation is performed through horizontal and vertical projection. The ANN is trained with individual Assamese handwritten character. From the line, character is over segmented using vertical projection. The trained ANN is used to validate character segmentation points.

In [100], the segmentation method based on Hough transformation and connected component analysis was used for line and word segmentation for unconstrained Indian script. The algorithm begins with applying connected component analysis on the document image. Further these components are divided into three categories based on component average height which is obtained through measuring average of all connected component. The subset1 components which are representing character belongs to average character height taking part in Hough transformation mapping and text lines are extracted. For word segmentation two different distance metrics, Euclidean distance and the convex-hull based metrics are used. The segmentation accuracy on handwritten text documents written in English, Marathi, Bangla, Kannad, Tamil and Malayalam scripts is tested with text documents containing minimum 21 and maximum 33 words. The average word segmentation accuracy of Indian script document is 76% and 90% for complex document and good documents respectively.

In [101], the painting technique was used for segmentation of Persian, Oriya and Bangla scripts. It is based on smearing technique to separate foreground and background portion of handwritten text documents enabling easy detection of text lines. After this process dilation is applied and whole line taken as single component. The algorithm also tested on ICDAR09 dataset and achieved 98.76% of recognition accuracy.

In [25], the two stage segmentation strategies based on morphological processing and projection profile was used for segmentation of handwritten Kannada document. The Kannada document is first applied morphological erosion and dilation operation. Initially connected component analysis is applied to document image to remove small components. After applying morphological processing, projection profile is used to detect lines. The algorithm is tested on 100 text documents containing 714 text lines and achieved 94.5% line segmentation. For word and character segmentation author used vertical projection profile and achieved 82.35% and 73.08% segmentation rate respectively.

In [102], the segmentation method based on connected component was used for reading bank cheque application written in Handwritten Devanagari script. The cheque image is converted to binary after applying smoothing and scaling. The preprocessed binary image processed for line extraction using clustering the connected component method. For extracting word from the line, vertical projection profile is used considering significant spacing between the valley points due to head line ("shirolekah"). To extract Devanagari characters from the word, vertical projection profile used to separate the base character using clear path between them. The segmentation rate reported is 98% word, 97% character segmentation and 100% for numeral segmentation accuracy if document properly written i.e. non-overlapping characters, proper connection of headline, proper spacing between words and character.

In [103], the vertical profile projection and horizontal profile projection was used to segment simple line and diagonal profile projection is used to segment skewed lines for Gurumukhi script text document segmentation. The algorithm requires user interaction for document type. The line segmentation method reported 93% of accuracy tested on 60 document images.

In [104], water reservoir based technique was used for segmentation of isolated and touching Gurumukhi characters. The Gurumukhi Isolated characters are well spaced and extracted after removing header line, while touching character extracted using the water reservoir whose height is greater than a threshold of 1/10th of the character, considered for further processing. The said method achieved 93.51% of segmentation accuracy for isolated and touching characters.

In [24], the connected component analysis and computation of variance method was used for segmentation of handwritten Kannada text lines. The algorithm applies connected component analysis to Kannada handwritten document after applying preprocessing steps like skew correction and noise removal. By ignoring small connected component and computing co-efficient variance of vertical coordinates through centroid, component is assigned to appropriate line. They have tested algorithm on 200 lines and achieved 96% of accuracy.

Based on literature review, majority of Indic scripts uses connected component based segmentation approach or its combination [24,100,105,106]. The advantage of using connected component is due to it extract whole character in line script without head line or word as single component in scripts with head line. Another approach which is widely used is piece-wise projection [23,25,103]. The main advantage of piece wise projection is due to its strip as within strip line skew do not affect the histogram. Also, we found that recognition approach followed by researchers for Indic script is mostly analytical [22,24,25,100,103,104,107].

2.2 LITERATURE REVIEW ON FEATURES AND CLASSIFIERS

The different types of features are used by different researchers for recognition of characters. Features are broadly classified into three categories: 1. Statistical features, 2. Structural features and 3. Global transformation and moments based [39,65]. Widely used features are based on projection [25,91,108], zoning [109-111] and structural [59,72,112]. Next section discuses features used by different researchers in their work.

In [108], profile counts, projection histogram and directional histogram was used as features. The profile counts and projection histogram features are calculated for whole image and normalize into 10 bins. The directional histogram feature is calculated by dividing character into 3 X 2 zones. They have achieved 87.8% recognition rate using metaclass classifier on the NIST SD19 database.

In [112], 280-dimension vector consist of horizontal histogram, vertical histogram, radial histogram, radial out-in profile and radial in-out profile was used as features. The radial histogram is calculated from the centre of the 32 X 32 normalized character images. Radial in-out profile is calculated based on recording first on pixel moving from centre pixel to periphery of the character. Similarly, out-in profile is calculated in from periphery to centre pixel. They have achieved rate of recognition 72.8% to 98.8% depending on database (NIST & GRUHD) and character category (digit, lower case, uppercase, mix) using k-means algorithm.

In [7], contour code based features was extracted from handwritten cursive English characters. The rate of change of slope is extracted as contour code. They have tested it on CEDAR dataset using feed forward neural network and achieved 84.93% on testing dataset.

In [113], zoning based technique was used, which calculates pixel density of object pixels as feature set. Isolated handwritten English numerals image divided into 4 X 4, 6 X 6 and 8 X 8 zones and extracted 116 density features from each zone. The classification was performed using k-NN technique and achieved 99.89% rate of recognition for dataset containing 10000 handwritten numerals.

In [114], zoning method for features extraction was used to recognize English alphabet numerals and two special symbols. The character image is normalized into 90 X 60 pixels size and divided it into 54 equal zones of size 10 X 10 pixels. The features are extracted from each zone along its diagonal lines which make up 19 sub features for each zone. These 19 sub features are averaged to form a single feature value and placed in the corresponding zone to make up 54 features. In addition to these 9 and 6 features are obtained by averaging the zone value row-wise and column-wise respectively resulting total 69 features. For classification feed forward back propagation neural network is used and achieved 98.50% with 69 features.

In [115], chain code histogram was used as features for Arabic character. The k-NN is used as classifiers and chain code histogram gives best results.

In [116], gradient was used as feature for English handwritten alphabet. The feature extraction technique, extract two types of features as global features and local features based on gradient. They obtained average rate of recognition 93.24% tested on 13000 samples collected from 100 persons.

In [117], principal component analysis (PCA) was used for recognition of Urdu characters. PCA transform isolated character images and placed into training dataset. This training image then projected onto Eigen space. They reported overall 96.2% of recognition rate for isolated handwritten character.

In [110], 40-point feature was extraction based on zoning for recognition of handwritten English alphabets. The character image then divided into 16 zones, further division of zone performed from corner of the image and middle of the image by combining 4 zones, 9 zones, and 16 zones. These makes total 40 zone features inputted to neural network. The rate of recognition achieved by MLP neural network is 83.84%.

In [111], diagonal feature extraction scheme was used for offline English isolated handwritten alphabet. A 100 X 100 character image is divided into 10 X 10 pixels zone. Their feature set consists of 300 values from which 100 features obtained from diagonal average, 100 features from row wise mean value, and 100 from Eigen value. To classify character a feed forward back propagation neural network having two hidden layers is used and reported 98.8% recognition accuracy.

In [52], uniform 5 X 5 zoning method was used for handwritten character recognition. A distance-based feature set was considered and a k-NN classifier, a feed-forward backpropagation neural network classifier and support vector machine classifiers were used. The approach, applied to the MNIST database, using 5000 training digits and 1000 testing digits, led to a recognition rate of 97.2% using the SVM classifier. When 12,000 training samples and 3000 testing samples of the ISI Bangla digit database were considered, the authors achieved a recognition rate of 95.47%.

In [79], novel feature extraction technique was used which is based on fuzzy-zoning and normalized vector distance measure. They reported 78.87% rate of recognition which is tested on dataset of 15,752 samples containing 44 basic Malayalam handwritten characters. The class-modular backpropagation neural network was used for classification together with a region-based fuzzy membership function.

In [106], the zonal moments features were used for recognition of Marathi Barakhadi for Marathi language. Character is divided into top region and middle region using the head line of Marathi character. The Hough transform is used to detect head line. The middle region further processed using vertical histogram for side of the consonant. The moment features then extracted from the both the region and classified using quadratic classifier for recognition of vowel and consonant part separately.

In [118], Zernike moment technique for feature extraction was used to recognize handwritten simple and compound character of Marathi. The Zernike moment has rotation invariant property. The characters are divided into three major categories based on global features. The global features are presence of vertical bar at right, vertical bar at middle and no bar. The vertical bar at right further divided into two categories based on bar’s connectivity with the character. The character image then divided into 9 zones and extracted end points and junctions for each zone. To extract Zernike moment based feature extraction image is segmented into 30 X 30 blocks. The SVM and K-NN used for classification and achieved 98.37% and 98.32% accuracy for basic and compound character respectively.

In [83], three stage feature extraction schemes were used for Malayalam handwritten character recognition. The group of characters formed based on geometric specification extracted in stage 1 and performed feature extraction for stage 2 based on group. In stage 3, constraints are imposed based on word formation and choose next matching character from the group. The rate of recognition is not reported by the authors.

As noted earlier, the amount of research in Gujarati handwritten character recognition is limited. In the next section, the literature review on Gujarati OCR is presented.

2.3 REVIEW ON HANDWRITTEN GUJARATI OCR

As discussed earlier in section 1.6 one can trace very little amount of work with respect to Gujarati handwritten OCR. Here are some of the significant contributions with respect to handwritten OCR in Gujarati.

In [33], work on isolated handwritten Gujarati numerals was presented by Baheti M. and Kale K. The preprocessing stages consist of binarization, bounding box segmentation, size normalization and skeletonization. The affine invariant moment derived for each of the numerals as feature extraction technique. Author used SVM, Gaussian distribution function, K-NN and PCA as classifiers. Amongst this classifier SVM shows higher recognition rate 92.28%. They experimented on 1600 images of handwritten numerals.

In [30], same author extended their work for noisy character. The binary cropped image is resize to 70 X 50 pixels and zone density features extracted for 10 X 10 pixels. The feed forward neural network was used as classifier and trained using 600 numerals and tested on 1200 images. The overall rate of recognition reported was 86% with highest recognition rate for numeral 9 in Gujarati.

Other efforts for recognition of Gujarati handwritten numeral was found in [45]. The horizontal, vertical and two diagonal projections are extracted as features from 16 X 16 normalized thinned image. A feed forward back propagation network is used for classification with 118 inputs, 60 hidden and 10 output neurons gives 80.5% recognition rate on 900 dataset.

The effort towards segmentation of Gujarati handwritten word discussed in [87]. The zone boundary for Gujarati word is identified through distance transform and then using connected component analysis each component is determined as basic character, modifiers, and connected modifiers. Zone boundary is used to dissect modifiers from the basic character. The said approach tested on 250 words gives accuracy 75%, 84% and 84% for correctly detection of upper zone, middle zone and lower zone respectively. The accuracy of correct extraction of basic character and modifiers is reported 82.4% and 60% for words without multiple component characters and with multiple component characters respectively.

The significant work on isolated alphabet of Gujarati characters is found in [27] with the success rate of 63.1% based on binary tree classifier and k-Nearest Neighbours (k-NN) for recognition of 45 Gujarati characters includes 10 Gujarati numerals and 35 consonants. The dataset for Gujarati handwritten characters was collected from 200 different writers. In preprocessing, adaptive histogram technique is used for contrast adjustment, two-dimensional adaptive Wiener filter used for noise removal, Otsu's thresholding method used for binarization and then character image is normalized to 40 X 40 pixels. They divided feature set into primary features and secondary features. The primary feature set consists of structural properties which divide the characters into subset for further recognition using secondary feature set with k-NN as classifier. The secondary feature consisting of 25 elements from averaging features based on 4 X 4 block, 28 elements moment based features, and 64 elements are centroid distance features. In [100], hybrid feature set was used to recognize Gujarati alphabets and numeral. The feature set consisting of aspect ratio, extent and zone density. To classify author used SVM with polynomial kernel (c=2), which gives 86.66% recognition rate.

In [85], authors worked on recognition of similar appearing characters. They reported that rate of recognition is affected by the shape of similar appearing characters. Authors used fuzzy K-NN classifier in pair with two additional Geometric and Wavelet features. The test data is collected from various sources like scanned pages, text books and newspapers and train data created through typing characters using different fonts. The rate of recognition with fuzzy K-NN is reported 100% for similar appearing printed Gujarati characters.

In [59], structural characteristics of Gujarati handwritten characters were identified which distinguish character from other characters. They have used decision table approach for classification of consonants.

In [119], stroke based features was extracted from thinned binary image for Gujarati character recognition. They have achieved 88% recognition rate using k-NN classifier for printed Gujarati numerals.

In [31], low level stroke features like endpoints, junction points, line segments, and curve segments was used for recognition of handwritten numeral written in Gujarati and Devanagari script. They have achieved 98.46% recognition rate for handwritten Gujarati numeral using k-NN classifier and results were further improved using radial bases function kernel in SVM classifier.

In [72], different features like aspect ratio, extent and zone density was used. The aspect ratio is calculated before converting character image into fix size. After that image is converted into 16 X 16 size and extent feature is extracted as total number of on pixel divided by area. Finally, 16 zone density features are extracted by dividing image into 4 X 4 sub images. The said 18 features give 86.66% rate of recognition using SVM classifier with polynomial kernel trained with 5000 Gujarati alphabets.

In [86], chain code based features was used for recognition of handwritten Gujarati numerals. They have obtained chain code using two approaches. In first approach chain code is obtained from the first pixel in horizontal direction. In second approach chain code is obtained from the pixel which is farthest from the centroid. After finding first pixel chain code is obtained. They achieved 96.37% and 95.62%

recognition rate for first and second approach respectively using feed forward back propagation network.

2.4 DISCUSSION

- It is noted that many standard datasets are available for comparison of rate of recognition for English, Chinese, Japanese, and Arabic which helps researcher to focus on improving HCR activities. While for Indian languages researchers create their own dataset.
- The solution proposed for different foreign languages and other Indian languages cannot be directly used for Gujarati character recognition because of the number of characters in character set and variable shape of characters.
- Although many researchers are working on English handwritten character recognition and till today searching for better accuracy like human.
- From the survey of literature, text line segmentation methods are classified into method that uses histogram, Hough transformation, smearing, component grouping, combination approach and others [20,22,26,28,68,69,109,120,121].
- Many researchers used preprocessing steps of document/character image which are suitable based on the choice of application and making images ready for feature extraction.
- The popular classifiers used for handwritten character recognition are neural network SVM, and tree classifiers.
- The selections of features used by many researchers for Indian languages are based on projection, profile, zoning, DCT, invariant moment and structural. Although, character recognition yet not achieved higher accuracy.
- In Gujarati one can find work on numerals recognition but very limited amount of work available for alphabet. Almost negligible amount of work done in case of Gujarati handwritten text recognition.