Pattern Recognition In Gujarati Character Detection – A Pragmatic Approach

Snehal Shukla[1], Dr. Purna Tanna[2]
[1, 2] FCAIT, GLS University

Abstract- *This paper presents the survey of work done on Gujarati characters recognition, success rate and limitations of it. Many researchers are working on different character recognition processes for Gujarati characters as well as numeric. They all have used different methods to recognize characters. We have listed some of them with their methods and results. So we can get idea about the current status of research work in this field, and one can decide their own way of research in the same field.*

I. INTRODUCTION

Pattern recognition is one of the core logical footmark in most actively growing field of OCR in linguistic digitalization. Digitalization modulate methodologies of preserving documents and also transfer of document. A huge literature and lots of government documents are there written by hands and needed to store in digital form to use it further. We can scan the document using scanner and can get image of it, but it will be only readable copy of it. If we want to modify it or update it we need to generate an OCR(Optical Character Recognition) for particular language.

India is a country with a large population of people with different culture and using verity of languages in routine. Approx 22 languages are spoken in India from which Gujarati is used by the people lived in Gujarat. Gujarati is reach with huge literature work and lots of document are available in Guajarati. In Gujarati language, there are 34 characters known as vyanjan, 13 vowels known as Swar, and 15 modifiers known as Matra as shown in fig 1.

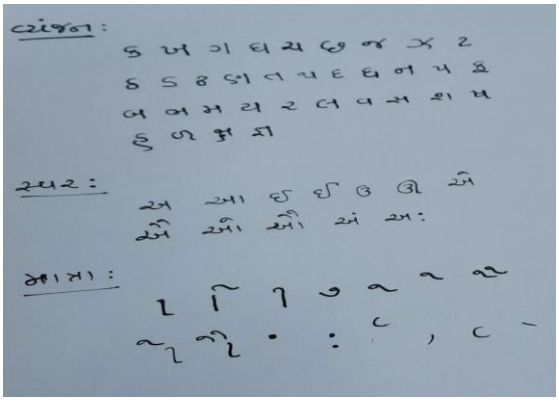

Fig.1

Gujarati is a language with large character set, number of combination of these characters are possible. It is a difficult task to recognize it, some tools are able to recognize printed characters but we found a research gap in recognition of handwritten gujarati characters.

In this paper we have analysed work done by other researchers and tried to find some conclusion for implementation of Gujarati Handwritten Character Recognition System.

II. RELATED WORK

First, we will start with need to convert handwritten documents to digital form. It is very useful for data entry of handwritten form, automatic conversion of prescription or script to digital form, automatic music notation reading etc. [1]. Gujarati handwritten documents will be first stored as an image, and we have to implement various processes to convert image into proper format so that we can use it for character recognition. The following are the steps to implement OCR in fig.2.

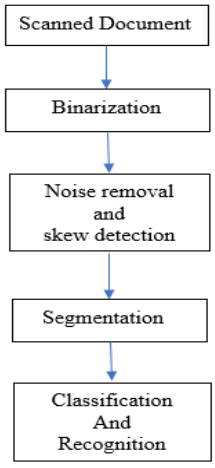

Fig.2

Binarization: It is a process to convert image into binary form (only black and white). K.Ntirogiannis et al[17] defines binarization as the process that segments the document image

into the text and background by removing any existing degradations.

There are many type of binarization techniques available like Gradient Based Thresholding, Niblack Method, Otsu Method, Nick Method, Bradley Method, Bersen Method, Local Adaptive Thresholding. Wan Azani Mustafa et al [15] have compared all the methods but using any method is not accurate for binarization. They have found Fuzzy C-Means algorithm with 97.02% accuracy for binarization[16].

Fu Chang [18] proposed a new method named Hadamard multiresolution analysis to get improvement in the output of binarization for OCR.

Thinning of image: It is a process of thinning of characters to a single pixel notation. Thinning is very useful process in image compression, biomedical image analysis, printed circuit board analysis, fingerprint analysis, etc.[19] and it is considered as an essential step for many OCR[20].Shaikh & shaikh[21] have listed the imitations of Lei-Hang's algorithm and proposed a modified algorithm for thinning process, which overcome the limitations of Lei-Hang's algorithm.

Maloo & Kale[2] defines steps for OCR as Binarization, Removal of noise, Thinning of image, Skew Detection and correction, Segmentation, Feature Extraction, Recognition using Classifiers. For binarization, they are using Threshold technique with level 0.4. Noise is wrong pixel value that are added to the image while scanning the image. Thinning process removes pixels those are not required for the characters; it will apply thinning repeatedly while all unwanted pixels removed.

Suthar, Goswami & Thakker[4] defines thinning as a crucial part before feature extraction, because to identify endpoint and starting of new character, if we have a thin line instead of thick line, it will be easier to detect. They have classified thinning process in two major class, 1) Serial thinning and 2) Parallel Thinning. In Serial Thinning, one point at a time will be processed and in parallel thinning all the points will be processed simultaneously. They define some essential properties for thinning as 1) thinning should be done till thickness of one pixel. 2) Connection of two characters or characters with modifiers should be preserved. 3)End point and junction point position should be preserved. 4) noise should be eliminated. 5) shape of the line and curve should be preserved. The authors have analyzed various Thinning algorithms like Hilditch, Stentiford, LW, ZW and ZS for normal dataset of printed text image as well as handwritten numerals and thick size of characters in dataset. They have defined 4 parameters that is Thinning Rate (TR), Number of Components, Noise sensitivity (NS), Endpoint accuracy for performance analysis. Using these parameters, they have found ZW algorithm is better in comparison of other because it gives 0.9827 performance in normal dataset and LW algorithm is better for thick character dataset with 0.8712 overall performance.

Skew Detection and Correction: When a document is scanned, some amount of tilt will be included, this tilt is known as skew angle. Skew angel should be removed from the image for correct input to recognition process.Arwa Al-Khatatneh et al[22] have compared (i) Projection Profile Analysis (ii) Hough Transform and (iii) Nearest Neighbour as skew detection techniques and given the strength and weaknesses of each algorithm. They found Nearest Neighbour method as the fastest method for skew detection. Chaudhary & S. Chaudhary [23] has implemented cross-corelation method for skew detection and found success with Monte Carlo technique for sampling. Gari et al [24] considers the skew angle detection as the most important component of OCR. They have implemented Harris Corner feature points and Hough Transform to detect the skew angle successfully for printed characters.

Shah, Patel & Maniar[4] have worked on Skew Detection and Correction of image. There are some important approaches for skew detection are Correlation, projection profile, hough transform and Linear Regression. Linear Regression is powerful technique for this. The authors have implemented Linear Regression for printed text as well as handwritten characters. It is providing 59.63% accuracy for printed document and 45.58% accuracy for handwritten documents. They suggest linear regression method is giving good result in 15-to-40-degree rotation, but less useful for minor rotation.

Chaudhary & Gulati [5] find the segmentation problem as the most important problem in the process of recognition of characters, because, segmentation improves the result of recognition rate. In Gujarati handwriting, implementation of segmentation is difficult due the different style of each writer. It can be affected by the type of pen used for writing. If gel pen is used, recognition is more difficult. To use ball point pen is recommended. The segmentation should be done first for the line, than for word and lastly for characters. The authors define main problems in 3 categories, 1) Problem in Line Segmentation. 2) Problem in word segmentation. 3) Problem in character segmentation. Problem in line segmentation includes problem like, overlapping of modifiers of two lines, zigzag of line and words and unusual spacing between lines. Problem in word segmentation is

unusual spacing between words of same line. Problem in character segmentation can be further divided as problem in upper region, problem in middle region, and problem in lower region. That can be further divided in problems like, unusual size of modifiers, touching of modifiers to the next character, touching of upper modifier with another modifier, determining presence of lower modifier, unusual spacing, touching of lower modifier to next consonant in middle region, touching+ of two consonants in middle region, touching of modifiers with consonant in middle region, touching of half character to full character, broken character, skewed characters. They have concluded that rate of successful segmentation will be decreased from printed document to handwritten documents.

Mendapara & Goswami[6] talks about identification of strokes, it is important part after character segmentation. Once we have segmented, we should identify strokes in the image for better feature extraction. For stroke identification and directional feature extraction of handwritten Gujarati numerals authors have used k-NN algorithm and achieved 88% accuracy.

Prasad & Kulkarni [7] have worked on feature extraction classification using Adaptive Neuro Fuzzy Classifier (ANFC) with Fuzzy Hedges (FH) for Gujarati character recognition. The authors say that using fuzzy hedges, we can find and remove noisy features and can select important features from the image. Feature Extraction can be using four features like 1) Gabor Phase XNOR pattern. 2) Pattern Descriptor. 3) Contour Direction Probability Distribution Function. 4) Autocorrelation. Using these features, they have achieved 58.78% accuracy for character recognition.

Hassan et al [8] have implemented MKL (Multiple Kernel Learning) for classification. They have implemented it with DAG framework and compared with 1-vs-1 framework using KNN classifier. They have implemented Shape Descriptor, Fringe Map Feature and Histogram of Oriented Feature (HOG) for classification of MKL. For binary MKL, Decision DAG framework is used. They have found best result using combination of Shape Descriptor and Fringe Map Feature.

Goswami & Prajapati[10] have listed many algorithms for classification like Pattern Matching, ANN, HMM, SVM, KNN etc. They have used Self Organizing Map based k-Nearest Neighbour(k-NN) classifier. They have got 82.36% of accuracy for selected characters in Gujarati printed character set. They have implemented this algorithm for various size and font of characters in different Gujarati news papers. Similar looking characters like tha, ya, pa, sha, gha, dha are difficult to classify, but unique characters are classified with their algorithm.

Desai[13] have used feed forward propagation neural network technique for recognition of Gujarati numerals. They have divided the process into 4 phases: digitization, Pre-processing, Profile vector and classification. In digitization, the captured image is converted into binary format. Pre-processing is divided into contrast correction, resizing, thinning and rotate training set. This proposed network give them 81.66% accuracy in recognition of gujarati handwritten numbers.

Patel & Desai[14] have defined three zones and two line to separate zones for any character: Lower zone, Middle Zone, Upper Zone and Mean Line, Base Line. They have described the importance of zone identification in the phase of segmentation. In Gujarati Language, most of the consonants are written in Middle zone, only the modifiers of them may written in Upper zone or Lower zone. So, if we identify the zone, we get help to identify character. They found self defined ANN algorithm useful for this segmented characters to identify, and got 60 % accuracy for 4 characters, that is (ga, na, sha, la).

Sr No	Authors	Input	Process	Algorithm	Accuracy
1	Maloo & Kale[2]	Gujarati Handwritten & Printed Characters	Binarization	Threshold with level 0.4	-
2	Suthar, Goswami & Thakker[3]	Gujarati Characters	Thinning Process	ZW for normal dataset LW for thick character dataset	0.9827 0.8712
3	Shah, Patel & Maniar[4]	Gujarati Numerals	Skew Detection	Linear Regression	59.63% for Printed Characters 45.58% for Handwritten Characters
4	Chaudhary & Gulati[5]	Gujarati Handwritten Characters	Segmentation	Defined Problems in Segmentation	-
5	Mendapara & Goswami[6]	Gujarati Numerals	Stroke Indentification	kNN	88%
7	Prasad & Kulkarni[7]	Gujarati Characters	Classification	ANFC(FH)	58.78%
8	Hassan et al[8]	Gujarati Characters	Classification	MKL using Shape Descriptor & Fringe Map MKL using Shape descriptor, Fringe Map & HOG	95.50% 92.43%
9	Patel & Desai[9]	Gujarati Characters	Classification	Binary Tree Classifier with kNN	63.1%
10	Goswami & Prajapati[10]	Gujarati Characters	Classification	Self Organizing Map based kNN	82.33%
11	Genzel et al[11]	Gujarati Characters	Classification	HMM	3.27% error rate
12	Desai[12]	Gujarati Characters	Classification	SVM with Polynomial Kernel	86.88%
13	Desai[13]	Gujarati Numerals	Classification	Forward Back Propagation Neural Network	81.66%
14	Patel & Desai[14]	Gujarati Handwritten Characters	Classification	Self Defined ANN	**60% with Modifiers for characters** ગ, ન, શ, લ

III. CONCLUSION

All the papers we have studied, have implemented OCR for Gujarati characters. Different result got by different authors. In gujarati characters recognition maximum 81% accuracy achieved. We are focusing on Analysis and recognition of Modifiers of handwritten gujarati characters. We have one paper that is concentrating on modifiers and achieved 60% accuracy with only four characters. So, we have a huge scope for analysing modifiers with all the gujarati characters. As we know that characters may fall under three zones, we are concentrating on the modifiers comes under upper and lower zone.

REFERENCES

[1] Scope of Handwriting Recognition in Indic Scripts by Sukhdeep Singh, IJRAR, Vol 6, Issue 1, Jan-Mar 2019.

[2] Gujarati Script Recognition: A Review by Mamta Maloo, Dr. K.V. Kale, IJCSI Vol 8, issue 4, no 1, July 2011.

[3] Empirical Study of Thinning Algorithms on Printed Gujarati Characters and Handwritten Numerals by Sanket Suthar, Mukesh Goswami, Amit Thakker, ERCICA 2014

[4] Skew Detection and Correction for Gujarati Printed and Handwritten Character using Linear Regression by Shah, Patel, Patel, Maniar, IJARCSSE, Vol. 1, Issue 1, Jan 2014.

[5] Segmentation Problems in Handwritten Gujarati Text by Chaudhary & Gulati, IJERT, Vol 3, Issue 1, Jan 2014.

[6] Stroke Identification in Gujarati Text using Directional Feature by Mahendra Mendapara & Mukesh Goswami

[7] Stroke Identification in Gujarati Text using Directional Feature by Prasad & Kulkarni, Springer-2014

[8] Use of MKL as Symbol Classifier for Gujarati Character Recognition by Ehtesham Hassan, Santanu Chaudhury, M Gopal, Jignesh Dholakia

[9] Extraction of Characters and Modifiers from Handwritten Gujarati Words by Chhaya Patel, Apurva Desai, International Journal of Computer Applications (0975 – 8887), Volume 73– No.3, July 2013

[10] Classification of Printed Gujarati Characters using SOM based K-Nearest Neighbor Classifier Mr. Mukesh M. Goswami, Mr. Harshad B. Prajapati, Mr. Vipul K. Dabhi, ICIIP 2011

[11] HMM-based Script Identification for OCR by Dmitriy Genzel, Ashok C. Popat, Remco Teunen, Yasuhisa Fujii, MOCR Aug,2013.

[12] Gujarati handwritten numeral optical character reorganization through neural network by Apurva Desai, Elsevier, Jan-2010

[13] Gujarati Handwritten Character Recognition Using Hybrid Method Based On Binary Tree-Classifier And K-Nearest Neighbour by Chhaya Patel, Apurva Desai, IJERT, vol-2, issue-6, June - 2013

[14] Binarization of Document Images: A Comprehensive Review by Wan Azani Mustafa, Mohamed Mydin M. Abdul Kader, ICoGeS, IOP Conf. Series, 2018

[15] Review of Different Binarization Approaches on Degraded Document Images by Wan Azani Mustafa, Hairy Aziz, Wan Khairunizam, Zunaidi Ibrahim, Shahriman AB, Zuradzman M. Razlan, ICASSDA, Aug,2018.

[16] A combined approach for the binarization of handwritten document images by K. Ntriogiannis, B. Gatos and I. Pratikakis, pp-3-15, Jan,2014.

[17] "Binarization of document images using Hadamard multiresolution analysis" by Fu Chang, Kung-Hao Liang, Tzu-Ming Tan, Proceeding of the 5th international conference on document analysis and recognition ICDAR'99, 06,Aug,2002.

[18] "A special skeletonization algorithm for cursive words," by Tal Stenherz, Nathan Intratot and Ehud Rivlin, Seventh International Workshop on Frontiers in Handwriting Recognition, pp.529-534, 2000.

[19] "Thinning methodologies -A comprehensive survey," by L. Lam, S.W. Lee, S.Y. Suen, IEEE Trans. Pattern Analysis and Machine Intelligence, pp. 869-885, 1992.

[20] "A Generalized Thinning Algorithm for Cursive and Non-Cursive Language Scripts" by Noor Ahmed Shaikh, Dr. Zubair A Shaikh, 2005 Pakistan Section Multitopic Conference, IEEE Xplore, March,2007.

[21] "A Review of Skew Detection Techniques for Document" by Arwa AL-Khatatneh, Sakinah Ali Pitchay, Musab Al-qudah, , 2015 17th UKSIM-AMSS International Conference on Modelling and Simulation, IEEE Xplore, September,2016.

[22] "Robust detection of skew in document images." By Chaudhuri, S. Chaudhuri, Image Processing, IEEE Transactions vol. 6, no. 2, pp. 344- 349. 1997.

[23] "Skew detection and Correction based on Hough Transform and Harris Corner" by Ahmed Gari, Ghislane Khaissidi, Mostafa Morbti, Driss Chenouni, Mounim El Yacoubi, , 2017 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), IEEE Xplore, May,2017.